Multicast Scheduling and Resource Allocation Algorithms for OFDMA-Based Systems: A Survey

Richard O. Afolabi, Student Member, IEEE, Aresh Dadlani, Student Member, IEEE, and Kiseon Kim, *Senior Member, IEEE*

***Abstract*—Multicasting is emerging as an enabling technology for multimedia transmissions over wireless networks to support several groups of users with flexible quality of service (QoS) requirements. Although multicast has huge potential to push the limits of next generation communication systems; it is however one of the most challenging issues currently being addressed. In this survey, we explain multicast group formation and various forms of group rate determination approaches. We also provide a systematic review of recent channel-aware multicast scheduling and resource allocation (MSRA) techniques proposed for downlink multicast services in OFDMA based systems. We study these enabling algorithms, evaluate their core characteristics, limitations and classify them using multidimensional matrix. We cohesively review the algorithms in terms of their throughput maximization, fairness considerations, performance complexities, multi-antenna support, optimality and simplifying assumptions. We discuss existing standards employing multicasting and further highlight some potential research opportunities in multicast systems.**

***Index Terms*—multicasting, resource allocation, multicast survey, scheduling, OFDMA, subcarrier allocation, power allocation, resource optimization, quality of service, multicast standards.**

I. Introduction

OVERWHELMING demands for high data rates and the need to support large number of users with flexible quality of service (QoS) requirements has led to an explosive surge in mobile and wireless communication systems development in recent years. These demands and requirements are anticipated to be more intense in the future as more military applications and commercial services become more prevalent. Of particular interest are certain applications which require transmission to selected groups of users that naturally lend themselves towards multicasting. For instance, geographic information updates such as traffic reports, local news, weather forecast, stock prices and location-based adverts. Multimedia entertainments such as IPTV, mobile TV, video conferencing, and other multimedia services, which currently account for one-third of mobile internet market, are some of the disruptive innovations that can be deployed using multicast technology [1]–[7]. Since there is no substitute for intelligent deployment and utilization of finite resources, hence, when multiple users within the same or adjacent cell require same content, multicasting allows such users to form groups and share allocated resources as illustrated in Fig. 1. The idea further maximizes spectral efficiency and minimizes transmission power consumption at the base station while also maximally utilizing the limited system resources [4]. This is in contrast to unicast transmissions where users cannot share resources and as many transmissions as number of users are required for full cell coverage.

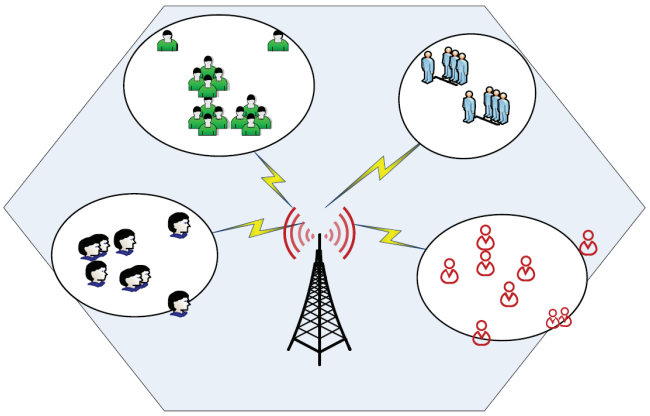

Fig. 1. Multicast system where multiple users requesting same service share allocated system resources. The users may not be in the same location.

Meanwhile, next generation communication systems must address challenges of multimedia broadcast due to wide variations of the wireless channel, high mobility of users and limited system resources. To resolve these challenges, combinations of multicasting together with orthogonal frequency division multiple access (OFDMA), multiple-input-multiple-output (MIMO) antenna scheme, scheduling, and dynamic radio resource allocation (DRA) have been particularly identified as spectrum efficient techniques to maximize spectral utilization, minimize transmission power consumption at the base station (BS) and provide better quality of experience (QoE) for users within the network. These technologies have been widely adopted as multimedia broadcast multicast services (MBMS) in few cellular standards such as IEEE802.16 (Fixed and Mobile WiMAX) and the 3GPP Long Term Evolution (LTE) to accommodate high speed mobility as well as support high rates for nomadic and mobile users [8].

The main idea of OFDMA is the distribution of the narrowband subcarriers among users depending on their channel

Manuscript submitted 10 May 2011; revised 05 August 2011 and 11 January 2012. This work was supported by the World-Class University Program through the National Research Foundation of Korea (R31-10026), and Grant K20901000004-09E0100-00410 funded by the Ministry of Education, Science, and Technology (MEST).

The authors are with the School of Information and Communication, Department of Nanobio Materials and Electronics, Gwangju Institute of Science and Technology, GIST, Gwangju, 500-712, South Korea (e-mail:**richoptix@gist.ac.kr**; dadlani@gist.ac.kr; kskim@gist.ac.kr)

Digital Object Identifier 10.1109/SURV.2012.013012.00074

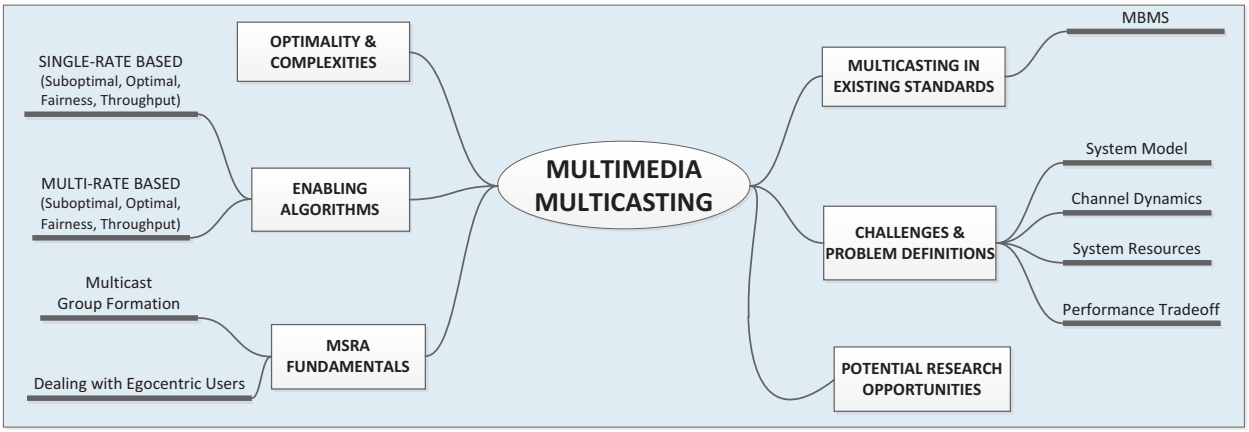

Fig. 2. Various aspects of channel-aware multicast scheduling and resource allocation (MSRA) as discussed in this paper.

characteristics [9], whereas, MIMO uses multiple antenna at both the transmitter and receiver to enable increased spectral efficiency for a given total transmit power by properly multiplexing parallel channels and taking advantage of antenna diversities. Similarly, scheduling and dynamic resource allocation establish management protocols to ensure fair and efficient exploitation of system resources. The transmission strength of OFDMA together with advanced antenna capabilities of MIMO allow more users to be packed into available resources in frequency and spatial domains. Combination of MIMO-OFDMA unique features has been reported to result in enhanced system total capacity [10].

Multicast scheduling and resource allocation (MSRA) is based on two types of multicast transmissions: Single-rate and multi-rate transmissions. In single-rate, the BS transmits to all users in each multicast group at the same rate irrespective of their non-uniform achievable capacities whereas in multi-rate, the BS transmits to each user in each multicast group at different rates based on what each user can handle. Until recently, single-rate scheme has been quite popular and widely accepted due to its implementation simplicity and low complexity. Multi-rate, on the other hand, has been receiving more attention lately because of necessity to achieve user throughput differentiation such that improved system spectral efficiency is attained.

MSRA is still confronted with various technical challenges. For example, in single-rate transmission, multicast services must be transmitted at a rate low enough for the least (worst or minimum) user to decode and high enough to maximally utilize system resources. Hence, the major problem is determining the most efficient single rate to transmit to each group without being insensitive to users with bad channel quality or unfair to users with high throughput potentials. Invariably, single-rate multicasting translates to trade-off between the transmission rate and system coverage.

In multi-rate transmission however, the problem is how to reduce the computational complexities, coding, and synchronization difficulties associated with transmission to multiple subgroups or individual group members. Based on these two types of multicast group rate determinations, scheduling, resource allocation and optimization can then be performed such that spectral efficiency is achieved, various network resources are optimally utilized without performance degradation and users' QoS requirements are satisfied given that they experience different channel fading dynamics.

While a huge plethora of literature exists on scheduling and dynamic resource allocation (DRA) in unicast multiuser OFDM systems as surveyed in [11], [12], works on multicast scheduling and resource allocation (MSRA) are just beginning to emerge in broadband wireless systems. Authors of [13] and [6] examined single-rate multiple multicast groups within a single cell while [14] and [15] investigated multiple multicasts with multi-rate transmissions. All these algorithms consider different performance metrics and constraints. Of particular challenge is the resulting optimization problem of multiple antenna complexities at both the BS and individual users. Specifically, [5] and [16] are among the few works investigating MIMO techniques in multicast. Hence, MSRA in wireless networks is currently a research area with many open issues.

A major goal of this survey article is to present a concise and insightful view of the current knowledge in several aspects of channel-aware MSRA algorithms and then provide succinct classifications of these algorithms as illustrated in Fig. 2. We start by introducing MSRA fundamentals and various group formation concepts in Section II. We discuss challenges associated with MSRA in Section III while in Section IV, we explain approaches in MSRA to address optimality and complexities. Main ideas, features and limitations of enabling algorithms and their various forms are studied in Section V. Multicasting features of some modern wireless standards are explained in Section VI. Finally, Section VII provides insight to some potential research opportunities and our conclusions are presented in Section VIII. Although, we do not claim absolute completeness of resource allocation algorithms in this study - because this would probably result in an heterogeneous list of scientific contributions - but extensive analysis has been provided.

II. Multimedia Multicast Fundamentals

In this section, we provide insight into various multicast group formation strategies proposed in the literature for single-rate and multi-rate multicast group transmission schemes. We also address potential anomaly behavior in multicast groups.

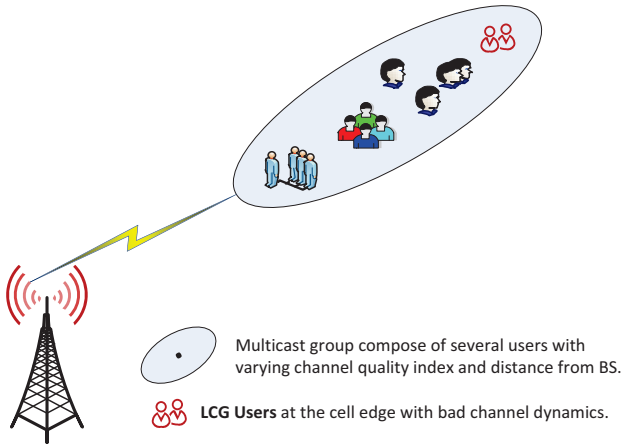

Fig. 3. Multicast group formation using single transmission rate. The single-rate can be fixed rate, average group throughput or rate of least capable user in the group.

A. Single-Rate Multicast Transmissions & Group Formation

Single rate transmission requires no special group formation except to determine a compromising transmission rate suitable for all users in the group as in Fig.3. To permit researchers to design and propose practical MSRA algorithms, three simple schemes have been adopted widely in the literature. First is a pre-defined fixed default rate [17], [18]. Second is adaptive selection and transmission at worst user's rate (i.e. user with least channel gain) [19] and finally, dynamic transmission using group average throughput [20]. In what follows, we discuss these schemes and their variants.

1) Pre-defined Fixed Rates: It has been argued that using a pre-defined fixed default group transmission rates for all multicast group is sufficient. In fact, existing communication systems such as CDMA2000 1xEV-DO networks use fixed data rate of 204.8Kbps for multicast transmission and equal resources are assigned to all users in cyclic round-robin fashion irrespective of their channel characteristics [17], [18]. This means, there is no priority consideration and system resources are evenly distributed. This easy approach is especially designed to favorably satisfy cell edge users who are expected to have low channel gains due to their bad channel quality resulting from their farthest distance from central BS.

However, in the likely event that we assume the fixed rate is always equal to the instantaneous achievable rates of the minimum users at the cell edge, then, using pre-defined fixed rate results in max-min fairness (see Section III-C2) since the resulting minimum user is given resource allocation priority to realize its maximum achievable rate. This rate is the worst rate because it assumes that there is always a user at the edge of the cell regardless whether such a user is actually present or not. Although, fixed rate approach is simple to implement with low complexity and also guarantees reliable multicast to users at cell edge; it is however undesirable since it puts severe restriction on achievable system throughput when users' channel differentiation is considered especially for those users close to the base station with good channel quality. Additionally, this scheme does not offer any utility maximization, hence, it is unresponsive towards intra-group and inter-group user fairness. *Intra*-group refers to the interaction and coexistence of multiple users within a *single* multicast group whereas *inter*-group refers to such *competitive* coexistence in *multiple* multicast where numerous groups compete among themselves for system resources.

2) Least Channel Gain (LCG) User Rate: A system is only as strong as its weakest link. So also is a single-rate multicast system based on LCG user. This scheme adaptively sets the group transmission rates to suit the user with the worst (minimum) channel quality [5]–[7], [19].While LCG scheme has spurred other approaches, the scheme itself is highly conservative and spectrally inefficient since users within the group (close to the BS) experiencing good channel gains are severely hindered from utilizing link adaptation to exploit their good channel gain. Besides, as the group size increases, the data capacity of the group becomes limited, because more users now share resources assigned to the group based on LCG user, consequently, capacity benefits of the multicast system diminishes as the number of users increases [21]. It is obvious that LCG scheme is a *pessimistic* special case of pre-defined fixed rates discussed in Section II-A1.

3) Average (AVG) Group Throughput: Another way to improve system capacity and exploit multiuser channel dynamics is to enable the BS transmits to each multicast group based on long-term moving average throughput of the group [13], [20], [22]. Group averaging technique has various forms. For instance, [22] orders users' instantaneous achievable throughput and selects the median throughput that can support half (50%) of all group member while in [20] and [13], authors develop models that allow the BS to select appropriate single data transmission rate based on the exponential moving average received throughput of each user inside the cell. Another important but yet rarely researched area is physical layer multicasting using multiple antenna where users' average signal-to-noise ratios (SNR) is used to determine the group's single transmission rate [23]–[25].

In [24], authors show that capacity maximization based on SNR averaging provides higher capacity than LCG scheme and further justifies that under certain scenarios, when users are mobile, SNR averaging corresponds to the LCG scheme. However, studies by Sun *et al*. [23] differ from [24] by showing that not only does the LCG scheme offer practical implementation benefits, but also satisfies the QoS of each user. Interestingly, results of [23] is only valid provided the optimized LCG SNR meets the threshold required for successful reception. Fundamentally, concept of group throughput or group SNR averaging premise on guaranteeing reliable transmission and successful decoding to half the user in the system. This overtly optimistic scheme invariably means certain packet loss is inevitable, especially for users far from the central BS who cannot cope with the high average group rates.

B. Multi-Rate Multicast Transmissions & Group Formation

Multi-rate multicast transmission emerges to address the sub-optimality that exists in single-rate transmission considering the intrinsic heterogeneous channel characteristics of wireless networks. This diversity, if not well addressed

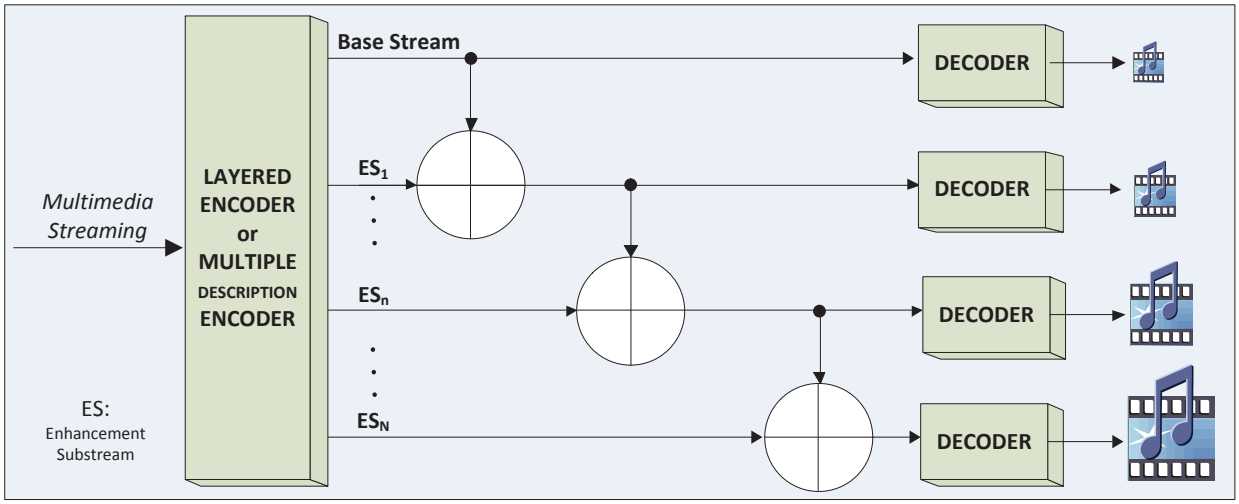

Fig. 4. Schematic of multi-rate information decomposition technique.

could have detrimental effects on multicast intra-group and inter-group transmissions. For example, single-rate intra-group unfairness may occur as a result of large differences in channel variations of users within a multicast group. On the other hand, inter-group unfairness may also result in networks with differentiated services where multicast groups requesting higher quality of services (QoS) would compete unfairly with other groups requesting simple best-effort services.

Hence, many researchers have concentrated on developing multi-rate multicast schemes allowing each user within a multicast group to receive multimedia traffic based on the handling capacities. Two techniques currently exist in literature for providing multi-rate multicast transmissions: One is *Information Decomposition Techniques* (IDT) (or *Stream Splitting*) [21], [26]–[29] which involves splitting high-rate multimedia contents into multiple streams of data where users subscribe to amount of data each can reliably receive. The other technique is *Multicast Subgroup Formation* (or *Group Splitting*) [20], [30], [31] which involves splitting and classifying multicast group into smaller sub-groups based on intra-group users' channel qualities. BS can then transmit to each group based on perception of what each group can accommodate. Recent studies such as [14] have developed variants of group splitting and also extended multi-rate multicast transmissions to multiple multicast groups with inter-group fairness considerations.

1) Information Decomposition Techniques - (Stream Splitting): Some authors have been actively investigating a physical layer information decomposition technique (IDT) or *multimedia stream splitting* which exploits user multichannel diversity [32], [33]. In IDT, information quality improves as users receive more substreams. Apparently, there is an exponential decrease in distortion experienced by users as received data rates increases. Two categories of IDT have been identified so far: *Multiple Description Coding* (MDC) [27]–[29], [34], [35] and *Hierarchical Layering* (HL) [21], [36], [37].

In MDC, multimedia data is split to multiple substreams and each subcarrier is allocated to one substream for downlink transmission to the multicast group. In both techniques, a minimal service - *base substream* - receivable by all users is defined and transmitted to all user. Afterwards, users with higher channel gain with potential for more throughput can also receive additional *enhancement substreams* (ES) to improve quality of the base substream. A schematic illustration of IDT is depicted in Fig. 4. The difference between MDC and HL is however in the order of reception of the substreams. In HL, substream ES_n must be successfully received before substream ES_{n+1} can be decoded whereas in MDC, all substreams have equal priority and any combinations of the received substreams can be decoded independently. MDC has enjoyed more attention from researchers because of its substream independence feature as it can potentially offer better performance than HL whose performance degrades when packets are not received in hierarchical order.

IDT multi-rate stream splitting and coding transmission without consideration for LCG users enable more significant performance than the conventional single-rate transmission scheme. In High Speed Downlink Packet Access (HSDPA) cellular network, IDT was shown to offer significant efficiency in utilization of allocated spectral resource [3]. However, the technique is reportedly not without significant computation complexity and signalling overhead. Although multi-rate IDT exploit users' channel variation and potentially perform better than single-rate, however, the assumption that multiple combinations of layers can be decoded by mapping the layers at the receiver to the original data still requires further investigation hopefully by adapting recent designs and advances in video coding technology.

2) Multicast Subgroup Formation (MSF) - (Group Splitting): Multicast subgroup formation (MSF) can be considered as an hybrid scheme to reduce the bottleneck effects of LCG users by combining the simplicity of single rate and higher capacity potential of IDT. In MSF, a multicast group is split into two or more subgroups and a single multicast transmission rate and coexistence mechanism are then defined for the subgroups. For instance, [35] divides a cell into two QoS regions. The BS transmits two data streams at different power levels according to the QoS definitions. Each stream corresponds to a different QoS region. High channel gain (HCG) users can receive both streams while the LCGs receive

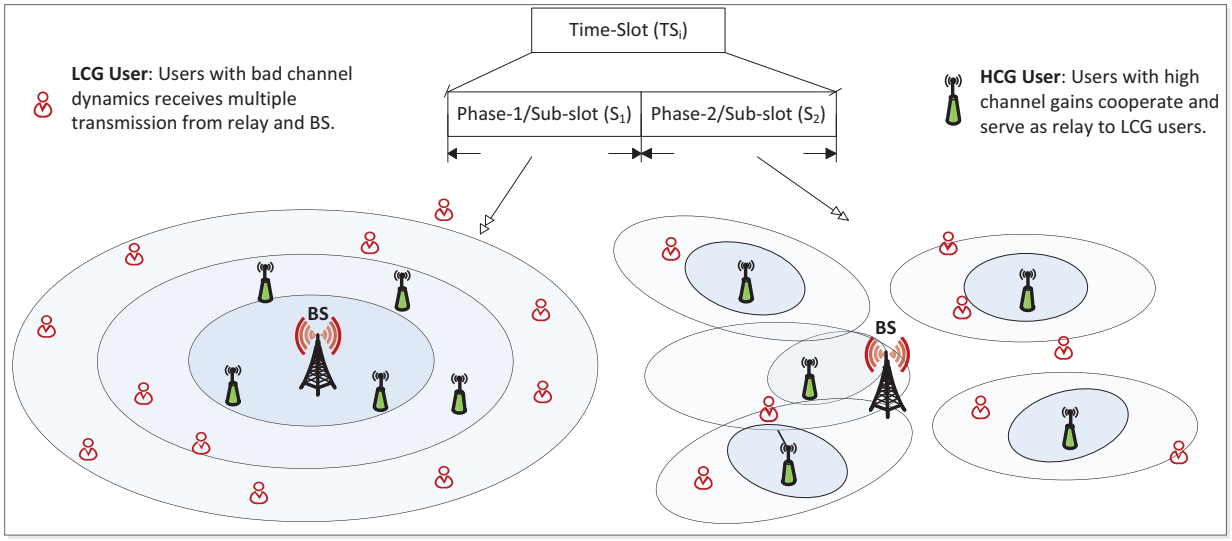

Fig. 5. Cooperative multicasting scheme. An example of multi-rate subgroup formation with cooperative data relay.

one stream. However, the scheme lacks precise coexistence definitions to guarantee reliable transmission to LCGs. Some coexistence ideas that have been proposed for use in MSF are itemized below:

- Subgroups + LCG users trade-off [20]
- Subgroups + subgroup resource sharing [30], [38]
- Subgroups + cooperative data relay [31], [39]

An example of multi-rate subgroup formation with cooperative data relay is cooperative multicasting scheme (CMS) [31] which exploits cooperative communication [40] as shown in Fig. 5. Hou *et al.* [31] proposed a time subslot-based cooperative multicast scheduling scheme for cellular networks where HCG users who reliably received the transmitted data in the time subslot 1, S_1, relay to the LCG users during S_2. Successful transmission in S_1 depends on the link quality between the BS and inter-group SNR rankings whereas in S_2, successful transmission to LCG users depends on intra-group channel diversity. Although CMS showed significant capacity improvement than the LCG based schemes; one potential drawback however is the energy inefficiency due to the over-concentration of resources on subslot S_2 retransmissions [31] (see Fig.8). This is apparent from the results showing that over 50% of HCGs within a group are expected to re-transmit in S_2 subslot. Techniques, such as nearest neighbor location-based service [39] and LCG maximum ratio combining [30] have been proposed to reduce the enormous power dissipation.

When *all* HCGs relay to the LCGs, the associated synchronization, estimation, and decoding complexities at the LCGs pose great overhead. Moreover, such general retransmission drains HCG relays by forcing them to expend enormous collective transmit power resource on LCGs. Additionally, more investigation is required on transmission reliability especially when the link quality between pairs of HCGs and LCGs deteriorates. This is possible due to shadowing or small-scale fading between the nodes which may ultimately result in intolerable error propagation and possible service denial for LCGs.

Although multiple subgroup formation offers higher capacities than single group, however rapid variation in users' channel dynamics is one major problem potentially making it impractical. This implies that group membership must change very often based on certain defined coexistence protocols. Associated with such rapid changes are heavy entry and initialization signaling overheads. Across the literature, MSF has been shown to be very effective in addressing multiuser channel and antenna diversities [31], [39].

Fig. 6 presents an intuitive schematic of the throughput of a multicast group having different channel-to-noise-ratio (CNR). Notice the sequence of rate increment for each user in each scheme. While IDT and CMS both have higher throughput potentials, they both also exhibit higher coding and synchronization complexities, demand for high computation power and require heavy re-transmission overhead. Since multicast services will almost certainly be required for future high data rate applications, it is highly critical that flexible and highly efficient algorithms be developed to provide high spectral efficiency under diverse QoS and traffic models.

C. Dealing With Egocentric Self-Serving Multicast Users

In designing reliable multicast transmission techniques, researchers have continuously made simplistic assumptions that multicast users are in their best, rational behavior and probably none will attempt to selfishly hoard resources. Hence, very few works exist considering anomaly behavior of intra-group multicast users. For example, in multi-rate subgroup with cooperative data relay [31], an inherent assumption is the unconditional, ever-willing to cooperate notion of selected relay nodes. Unfortunately, relay nodes have been shown to be egocentric and reluctant to contribute or forward data and expend extra power on other nodes [41]–[43]. This pattern of behavior is typically expected of mobile nodes with limited battery power self-perception since relaying data is energy-intensive.

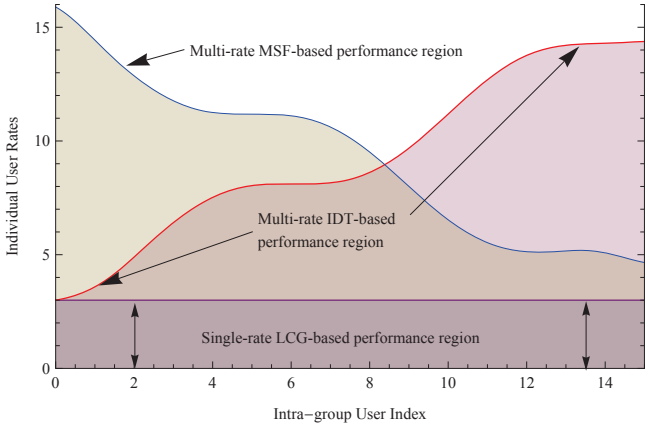

Fig. 6. Schematic illustration of single and multi-rate schemes: (a.) Single-rate group where threshold user with worst CNR bottlenecks other users' rate. (b.) IDT multi-rate stream splitting scheme where LCG rate first serves as base stream; other users receive enhancements streams ES_n proportional to their CNRs. (c.) CMS multi-rate group splitting scheme where HCGs are first served, then relayed to others low CNR users.

Some relevant questions to ask are:

- Are multicast nodes compelled to join multicast groups and/or forward received data?
- What benefits do such users stand to gain knowing that certain unknown users with unfavorable channel quality can constitute transmission bottleneck which potentially limits their achievable rates or cause them to incur expensive energy cost?

Answers to the aforementioned questions from users' perspective would probably negate any logical reasons for users to join multicast groups. Consequently, incentive mechanisms are required to motivate users' cooperation and discourage self-centeredness in multicast systems.

Three categories of incentive mechanism classes have been identified to stimulate user cooperation: First is *payment-based mechanism* [44], [45] where nodes that get services should be charged and those who provide support should be remunerated. Second, *reputation-based mechanisms* [46]–[49] where nodes monitor each other's behavior and cooperate with those who maintain good reputation. Lastly is the *punishment-based mechanisms* [50], [51] where non-cooperative nodes are punished by employing some punitive strategy.

Unfortunately, majority of these works target unicast systems and they cannot be directly applied to multicast. Perhaps, one of the few works dealing with this problem in multicast system proposed a game theoretic, punishment-based mechanism where nodes decrease their transmission power or marginalize erring nodes when misbehavior is detected [52].

Anomaly behavior evaluation in multicast self-configuring cognitive radio system is particularly a viable research opportunity to explore as many unresolved problems still exist. Research in this direction is particularly crucial because of the potential limiting impacts self-serving users may have on total network throughput. Possible solutions could involve evaluation, adaptation and extension of some existing algorithms already designed for unicast systems.

III. MSRA Challenges

In this section, we briefly describe various challenges associated with MSRA algorithm design. Then we present a multicast system scenario, its associated constraints and related resource optimization solutions.

A. Wireless Channel Dynamism

Signal impairment is a major problem resulting from rapid wireless channel variations, multipath propagation and fading of transmitted signals. Likewise, the mobility of users and attributes of the surrounding terrain make the wireless link to vary considerably in frequency, time and space for all users especially in urban areas where multicast is especially beneficial. This channel diversity of users can be adapted to assign subcarriers, modulation, coding rate and transmit power to users based on their instantaneous channel experiences. Hence, we can ensure that the most efficient transmission mode is always employed on each subcarrier regardless of the wireless channel quality. For example, in WiMAX, 64QAM $\frac{2}{3}$ coding rate having high spectral efficiency may be used when the mobile nodes are close to the BS and the link is good but BPSK $\frac{1}{2}$ coding rate having poor spectral efficiency but good bit-error-rate (BER) is used when the MS is far from BS and link is bad. In multicast systems, the use of aggressive modulation is complicated due to channel disparity of users, hence, little progress has been achieved in this avenue.

B. Channel-Aware Resource Allocation

Adaptive resource allocation is based on few notions: firstly, it is assumed that nodes can perfectly estimate and feedback their channel state information (CSI) to the BS. Secondly, CSI is always available to the BS before the commencement of each transmission. Thirdly, channel is slow-fading, meaning that the channel condition does not change during each OFDM symbol transmission block to avoid allocating resources based on obsolete CSI. While these assumptions are required to investigate system performance; they however do not offer realistic view of practical wireless systems. Practically, perfect CSI is hardly ever available at the base station due to channel prediction error, quantization error, feedback overhead, and channel feedback delay which is due to variation of the wireless link after estimation [15]. This delay often nullifies the validity of the estimates and degrades system performance.

The need for perfect CSI is even more apparent in channel-aware MSRA algorithms since the BS often utilizes knowledge of the channel condition for transmissions. BS would require frequent updates of the channel information for efficient resource allocation and optimal utilization. This frequent feedback requirement imposes significant load and complexity on the system since multiple multicast groups within a single cell may use large number of aggregate subcarriers. Therefore, some authors have studied channel estimation techniques that reduce feedback overhead [53], [54]. More specifically, [53] investigated the sufficiency of partial CSI in maintaining certain performance level and [54] considers the impact of delayed feedback channel on system throughput and shows that predictive coding can be used to mitigate effect of outdated CSI.

C. MSRA Performance Metrics & Tradeoffs

The principle of multicast communication hinges on how to achieve intra/inter multicast group rate balance (or fairness) and compromise. There are two major conflicting fairness criteria in the literature depending on various perspectives. First is equality based fairness where all users expect equal rates or resource shares. Second is proportional fairness where each user receives allocation based on their potential capabilities. Our work tends towards the latter definition. Since fairness and throughput maximization have always been two conflicting issues of concern in resource allocation problems, therefore, trade-off (compromise) or proportionality is always required to obtain good performance. Although these issues have enjoyed tremendous research attention for conventional unicast systems in recent years, more investigation is still required for multicast systems. In this section, we classify MSRA algorithms into three main categories depending on their features.

1) Strict Throughput Maximization (STM): Without consideration for fairness, STM is often utilized in multiple multicast resource allocation problems where inter-group competitive coexistence must be well managed to achieve optimal system spectral efficiency. STM is an overtly optimistic approach which is totally inapplicable in intra-group because it selects rate of user with highest channel gain as the group's transmission rate which undoubtedly would result in absolute intra-group resource starvation. On the other hand, STM has been shown to attain significant capacity gains for inter-group resource allocation because it allocates the best resource in time, frequency, and spatial domains to groups with the best potential to maximize total system capacity [19], [55], [56]. However, the gain comes at the expense of groups composed of sizeable number of users experiencing poor channel quality.

2) Max-Min Fairness (MMF): In multicast-enabled systems, MMF attempts to rectify fairness deprivation in STM by giving priority to minimum users/groups to realize their maximum achievable rates. The intra-group single-rate schemes explained in Subsection II-A are variants of MMFs depending on different threshold-rate definitions. For instance, in [13], [20] the threshold is defined as average (AVG) group throughput. Both schemes then traded-off LCG users and full system coverage. To achieve higher system capacities, the system is then optimized in favor of the AVG users. An improvement to both schemes could be to give the LCGs higher priority in the next time slots when their channel gains might have improved. Consequently, LCGs would experience temporary service failure instead of denial-of-service as in the case when they are totally given up. In MMF-based MSRA algorithms, each iteration maximizes the threshold-rate by allocating resources to users to achieve their highest possible rates until pareto optimality is attained - this is a state at which there is no other way to improve allocation of system resources to threshold user *without* decreasing allocation of other users. Hence, MMF-based MSRA algorithm is a pessimistic approach to provide guaranteed reliable multicast transmissions to all users.

3) Proportional Throughput with Fairness (PTF): Absolute fairness leads to drastic reduction in aggregate throughput. Strict throughput results in zero tolerance for weak multicast groups. However, proportional fairness is a compromise-based approach which attempts to simultaneously balance group aggregate throughput while preventing resource starvation and providing fair QoS to all groups. In [57], an elegant tradeoff factor approach is employed to manipulate proportionality between fairness and total system capacity. Results show high capacity gain with good fairness performance, however, unicast system was considered. A more related study conducted in [13], [36] considers multiple multicast groups scheduling in TDMA-based cellular data networks and proposed two algorithms optimized for intra-group PTF and inter-group PTF. In the inter-group PTF scheme, the BS dynamically selects multicast group such that the summation of $log(T_g)$ for all multicast group is maximized, where $log(T_g)$ is the group throughput for multicast group g. These algorithms are particularly interesting if we note that a system may achieve high spectral utilization, yet, a number of users still experience resource starvation. In such cases, the efficiency of the system results from users with good channel quality.

D. Multicast System Model

To implement channel-aware MSRA, users' CSI are assumed to be known at the BS. CSI is estimated at each user node and sent to the resource allocation block in the BS using feedback path. Alternatively, CSI can be estimated through the uplink of the user node in a time division duplex (TDD) system. As depicted in Fig. 7, the BS utilizes the CSI thus, obtained to assign a set of subcarriers to each user. It also determines number of bits to form an OFDM symbol, modulation scheme and amount of power to transmit on each subcarrier. When each OFDM symbol is transmitted, the subcarrier and bit allocation information are also sent along to the receivers through the control channel, thereby enabling the receivers to make informed decision about bits decoding and extraction from the sets of subcarriers assigned to the multicast groups.

Assume an OFDMA-based system with κ users on N subcarriers receiving multicast downlink traffic flows from a central BS having G multicast groups. Sets of users receiving the gth traffic flow can be represented as K_g, whereas number of users in a multicast group is $|K_g|$[1]. We denote total number of users in the system as $\kappa = \sum_{g=1}^{G} |K_g|$. Each group has fixed or variable number of users with different channel characteristics who may be co-located or differently located. The wireless channel is a frequency selective Rayleigh fading channel and the noise power of every subcarrier is assumed to be unity for simplicity. Each subcarrier has equal bandwidth size of $B_W = \frac{W}{N}$, where W is the total bandwidth of the system. For simplicity, we consider an MSRA LCG-based single-rate multi-multicast system where each multicast group rate is limited by the least-capable user. If $\min_{k \in K_g} |h_{k,n}|$ is the channel coefficient of minimum user k in group g on subcarrier n, N_0 is the white noise single-sided power spectral density on each subcarrier, then the frequency channel-to-noise-ratio (CNR) of group g on subcarrier n is $\hbar_{g,n} = \frac{\min_{k \in K_g} |h_{k,n}|^2}{N_0 B_W}$.

[1]Notice that multicast group results when $|K_g| > 1$. When $|K_g| = 1$, we have conventional unicast group.

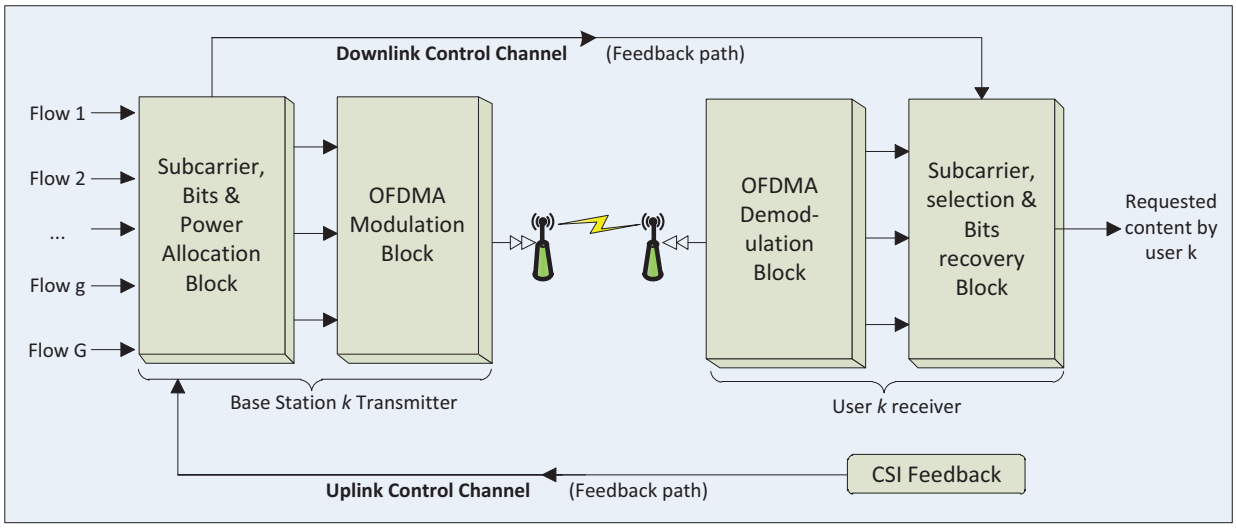

Fig. 7. Block diagram of multicast resource allocation block in OFDMA system. Subcarriers, bits and transmit power allocation are decided using resident MSRA algorithms. The CSI feedback block provides channel statistics from receivers to the base station.

Note that $\hbar_{g,n}$ captures the path-loss, fading, and noise of all the multicast users. Fundamentally, throughput experienced by each user depends on the number of users in each group and the differences in channel quality of each user. Therefore, multicast group transmission rate $R_{g,n}$ on subcarrier n is then given as:

$$R_{g,n} = \frac{1}{N}\log_2\left(1 + p_n \hbar_{g,n}\right), \tag{1}$$

where p_n denotes the amount of transmit power allocation on subcarrier n. Moreover, since more than one user can be allocated to a single subcarrier, we define a subcarrier allocation index, $\delta_{g,n}$ showing if a flow received by certain group occupies the n-th subcarrier or not. Note that here, intergroup subcarrier sharing is not permitted. Hence,

$$\delta_{g,n} = \begin{cases} 1, & \text{if subcarrier } n \text{ is allocated to group } g. \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The total throughput of a particular group g on all N subcarriers is then given as R_g in eqn (3).

$$R_g = \sum_{n=1}^{N} \frac{|K_g|}{N} \delta_{g,n} \log_2\left(1 + p_n \hbar_{g,n}\right), \qquad \forall g, \tag{3}$$

The underlying MSRA problem is basically to determine the most efficient way to allocate system resources, the optimal rate the BS should transmit to groups, which subcarrier(s) should be assigned to which group, and the required power for transmission on each subcarrier of each group. Then, the resulting optimization problem to improve total system capacity C_T becomes a non-convex, mixed-integer, non-linear maximization problems which is NP-Hard as shown in eqn. (4)-(7). NP-hard (Non-deterministic Polynomial-Time) problems are classes of problems for which no efficient solution exists [58], [59]. Results of the optimization problems give set of optimal subcarriers and power allocations maximizing the system capacity.

$$\max_{\delta_{g,n}, p_n} C_T = \sum_{g=1}^{G} R_g \qquad \forall n = 1, 2, ..., N, \tag{4}$$

subject to:

$$\sum_{n=1}^{N} p_n \leq P_{Total}, \quad \& \quad p_n \geq 0, \tag{5}$$

$$\sum_{g=1}^{G} \delta_{g,n} = 1 \qquad \forall n, \tag{6}$$

$$\delta_{g,n} \in \{0, 1\} \qquad \forall g, \forall n, \tag{7}$$

Equations (5) & (6) show that the total transmit power on all subcarriers cannot be greater than the system transmit power P_{Total} available at the BS, where eqn. (7) is the integer constraint defined in eqn. (2). Note that the complexity and hardness of this global optimization problem is due to the integer constraint and it becomes more difficult with increase in number of users and subcarriers. Since computation complexities increase with number of individual subcarriers to be allocated, it may be potentially helpful to allocate the subcarriers in *chunks* or blocks to reduce complexity. In [60], [61] and references therein, it was shown that chunk-based contiguous subcarrier allocation method based on SNR or BER constraints can effectively mitigate complexities and overheads. However, as expected, one major drawback of this approach is how to combat frequency selective fading on some subcarriers within the chunk which may hamper the possible benefits of chunk allocation.

In general, the cross-layer resource allocation and optimization problems [62] to meet the QoS requirements for all services requested by multicast users, maximize system throughput, maintain user fairness, minimize user and base station transmit power while considering channel characteristics of each user in multi-antenna OFDMA system is extremely challenging and sophisticated techniques with low complexities are still required.

TABLE I
OPTIMAL & APPROXIMATE MSRA ALGORITHMS COMPLEXITIES

Implementation	IDT-Based	LCG-Based
Global Optimal	$O(NN_d \sum_{g=1}^{G} 2^{\lvert K_g \rvert})$ [29]	$O(G^N \cdot N^{p'})$
Approximation	$O(NN_d\kappa)$ [29]	$O(\kappa N^2)$

IV. OPTIMALITY & COMPLEXITIES

Multicast problems fundamentally result in reduced system capacity due to the dependency on least channel gain user. Therefore, most existing work generally consider optimization problem with system capacity (C_T) maximization under various constraints including QoS, BER, delay tolerance, number of subcarriers, available BS power, etc. Results of the optimization solution gives set of optimal resource allocation maximizing the system capacity. In [19], [28], [29] , optimal solution to the objective function of the optimization problem have been investigated and shown to be highly desirable for different system resources. However, it requires joint allocation of multiple resources (e.g. subcarriers and power allocation) resulting in NP-hard, non-convex constrained optimization problem. Available solutions to this problem (e.g. exhaustive search, approximations) often require very complex and time-consuming computations.

To solve the problem, the difficult integer constraint $\delta_{g,n} \in \{0,1\}$ can be relaxed to assume continuous values $\delta_{g,n} \in C(0,1)$, maximization problem converted to minimization problem and the constraints of the maximization problem transformed to sets of linear equations and inequalities to obtain a relaxed convex feasible problems solvable by using linear programming (LP) [63] depending on the problem formulation. The resulting relaxation transforms the NP-hard mixed-integer (MIP) optimization problem into tractable convex LP problems whose solutions provide a bound - i.e. near-optimal approximations - on the original optimization problem. A probing question then is: "what is the performance gap between the exact optimal and its relaxed approximation?" This is a classical optimization discussion and myriad of papers address this issue for unicast systems. What is vital however is, at rare situations, relaxed optimal solution may have integral values $\delta_{g,n} = 0$ or $\delta_{g,n} = 1$ which then coincides with the exact MIP optimal solution.

Even with relaxation, computational complexity of the relaxed solution is still prohibitively high at the base station. Understandably, if the number of subcarriers and users within a cell are quite small, then, optimal solution with exponential complexity may suffice. Unfortunately, this is not the case because the number of constraints and variables increase significantly with the number of users and subcarriers, thus, making computational complexity unavoidably high. This is even worse when we consider that the optimal solution should be recomputed in response to each channel fluctuations. This is not realizable in practical systems like WiMAX where allocation must be achieved real-time within 5-10ms. Table I shows the computation complexities of optimal and near-optimal approximate solution of two algorithms[2].

[2]Number of possible subcarrier allocation is denoted as G^N while $N^{p'}$ is the number of searches for optimal power allocation over the N subcarriers such that system capacity is maximized.

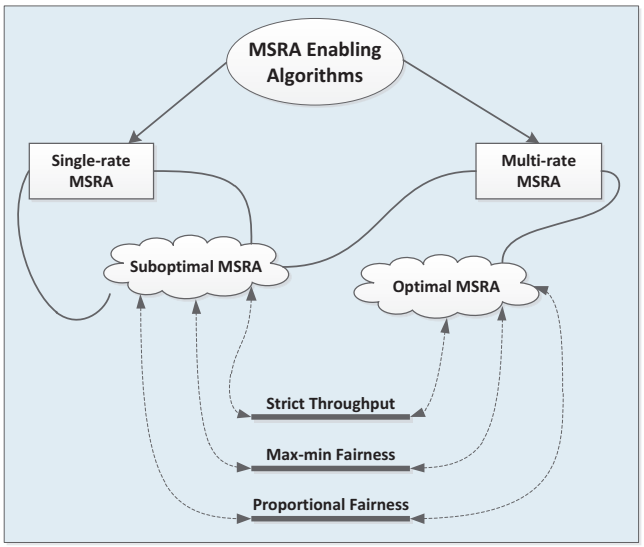

Fig. 8. Organization of the MSRA algorithms.

Consequent to these complexity problems, the MSRA optimization problems are often decomposed into independent, two-step operations and suboptimal/heuristic algorithms with lower complexities are often designed to solve the problems [5]–[7], [19], [36]. In numerous studies, heuristic algorithms have been shown to provide feasible solutions, however, the quality of the solutions or gap from the exact (and approximate) optimal solution remains uncertain. One of the few works performing such comparative analysis [27] reported a 5% performance gap between the relaxed optimal approximation and the proposed heuristic algorithm. Authors of [14] also examine the lower and upper performance bounds of the LCG-based and MDC-variant of IDT multicast transmission and proposed a low-complexity algorithm using Newton's Method [63] by exploiting the nonzero linear function power allocation constraints. Although, comparative study of exact optimal solutions would involve intensive computation power but it is nonetheless essential to investigate how these heuristic algorithms stack up with exact global solutions. Authors of [29] provide required work in this direction.

V. MSRA ENABLING ALGORITHMS

Various MSRA algorithms and optimization techniques with different objectives and constraints have been proposed for the downlink of OFDMA systems. We present details of some of these proposals in this section. Specifically, we evaluate their core characteristics, limitations, simplifying assumptions and performance complexities. Using multidimensional matrices, we cohesively classify them in terms of their throughput maximization, fairness considerations, multi-antenna support, optimality and sub-optimality. Algorithms discussed in this section are divided into different categories as shown in Fig. 8 as well as Tables II, III and IV. In the following, we discuss a few of the classifications.

TABLE II
SUMMARY OF **SUBOPTIMAL** MSRA ALGORITHMS FOR SINGLE RATE GROUP TRANSMISSIONS (FIXED, LCG, AVG, & LCG VARIANTS)

	Algorithms & Features	Advantages	Limitations
Strict Throughput	**SPA** [38]: LCG-variant. Subcarrier-only allocation for situation where user k has good link on subcarrier n when other K_{g-1} users have bad channel dynamics on subcarrier n. Users receive one pair of subcarrier. Min of max rates on paired subcarrier is used as intra-group rate then the group with max rate is scheduled for resource allocation.	Transmit redundancy on multiple subcarrier can possibly ensure robustness and make LCG user more competitive.	It lowers number of uniquely allocateable subcarriers by 50%. Also, lack of subcarrier pairing criterion can degrade system performance.
	LcSPA [19]: LCG-based Low-complexity, subcarrier & power allocation scheme that is largely applicable in multiple multicast single rate transmission. Adopts two-step approximation scheme. Two other variants of algorithm (L-LCSPA & H-LcSPA) are designed for low and high SNR regions respectively.	Throughput maximization. Reduction in total computational complexity due to linearity of the algorithms.	Lack of fairness. Can potentially starve out groups with poor channel gain or fewer numbers of users.
	MIMO-OFDMA MSRA [5]: Multiuser multiantenna version of LcSPA [19] using spatial multiplexing in single multicast. Extended in [16], for multi-multicast with MIMO precoding. [5] Adopts two-step approximation scheme as LcSPA [19] while [5] included third steps for computing optimal precoding vector.	Same as LcSPA [19] in addition to addressing MIMO complexity issues.	Same limitations as LcSPA [19]
Proportional Fairness	**Genetic Algorithm** [6]: Multicast Greedy Algorithm that iteratively allocates subchannel and transmit power. two-step suboptimal approach designed for multiple multicasts with proportional fairness considerations.	Uses compensation scheme to enforce proportional fairness.	-
MaxMin Fairness	**PSA** [64]: LCG-variant. two-step algorithm, divides each group to two subgroups based on low/high channel gains. Low subgroup users cooperatively share subcarriers with high gain subgroup. Attempts to proffer user fairness such that multicast group with severe LCG user can still receive some low rate data instead of total denial of service.	Potentially effective in best effort traffic multi-multicast system where groups containing users with severely low gains can be totally starved out of resources.	In practice, it is unlikely that the low rate data received by the low gain subgroups provide any satisfactory QoS for the much needed multimedia multicast traffic.
	CARA [56]: LCG-based Convex-optimization Aided Resource Allocation two-step suboptimal. Guaranteed QoS support.	Ensure QoS satisfaction for LCG user by adding minimum rate constraints to the optimization problem.	May result in resource starvation for whole multicast group in low SNR region where LCG user cannot meet QoS requirements.
	SPA [38]: Subcarrier Pair Allocation. Single-rate LCG-variant that attempt to make LCG more competitive in multi-multicast by using subcarrier pairing. Single step suboptimal. Introduced subcarrier pairing and transmission redundancy to evaluate user rates.	Ensures robustness. Computation complexity reduces by 50% of other schemes since no power allocation is required.	Reduction in number of usable subcarriers and lack of criterion for subcarrier pairing can potentially degrade system capacity and performance.
	Fair-SA [7]: LCG-based, Fair Subcarrier Allocation for preventing greedy resource utilization in single-rate, multi-multicast system. two-step suboptimal approach. Additional scalable fairness constraints added to the multicast optimization problem.	Implements group fairness of access to system resource by guaranteeing flexible minimum number of subcarriers to all groups.	Scheme assumes residual subcarriers are always available to compensate LCG user. Also, minimum user becomes bottleneck on group and system throughput.
	Averaging Schemes [13], [20], [22], [24]: Utilize SNR averaging, group throughput moving average, median instantaneous throughput to determine the minimum threshold or benchmark for resource allocation. Each iteration attempts to maximize the defined benchmark rate for each group. Guarantee reliable transmission and successful decoding to half the user in the system.	They are characterized by increased capacity and enhanced spectral efficiency for users capable of decoding the transmitted data. It also potentially increases inter-group competitiveness.	Results in high packet drop for high number of users especially LCGs which might have been dropped due to potentially low rates.

TABLE III
SUMMARY OF **OPTIMAL** MSRA ALGORITHMS FOR MULTI RATE GROUP TRANSMISSIONS (MDC, CMS & VARIANTS)

	Algorithms & Features	Advantages	Limitations
Strict Throughput	-	-	-
	-	-	-
Proportional Fairness	**MDC-IBL** [28]: Maximum Throughput & Proportional Fairness. IDT-based single multicast. Developed optimal (not in all cases) iterative bit loading scheme for weighted sum rate optimization.	Each user's capacity is a reflection of own channel quality.	High coding, decoding complexity and heavy transmission overhead.
	MDC-Duality [29]: IDT-based scheme with discrete modulation extending [28] to asymptotically global optimal solution for multi-multicast system. Provides global optimal maximum weighted sum rate using Lagrangian dual.	Performance significantly better than LCG-based and MDC-IBL [28]. Also claims algorithm has low complexity and fast convergence.	-
MaxMin Fairness	-	-	-
	-	-	-

TABLE IV
SUMMARY OF **SUBOPTIMAL** MSRA ALGORITHMS FOR MULTI RATE GROUP TRANSMISSIONS (MDC, CMS & VARIANTS)

	Algorithms & Features	Advantages	Limitations
Strict Throughput	-	-	-
	-	-	-
Proportional Fairness	**CMS** [31]: channel-aware two-phase subgroup where BS transmits to all multicast group at high rates during subslot S_1. Receiver at S_1 serve as relay nodes to LCG subgroup at S_2. Cooperative scheme exploiting intra-group user channel diversities to cooperatively retransmit to LCGs. Addresses the subgroup formation criteria lacking in [35].	Significant system throughput improvement than LCG scheme and ensures proportional rate fairness based on each user's channel gains using modified normalized signal-to-noise ratio ranking between the groups.	Highly energy inefficient due to the over-concentration of resources on subslot S_2 with 50% of users retransmitting as relay node.
	E-CMS [39]: Uses channel information and nearest neighbor discovery LBS to select relay nodes such that cumulative cooperative energy consumption and number of HCG relays are minimized. It has same features as [31].	Increases total system throughput and established variable number of relay nodes depending on cell coverage ratio. This potentially reduces network power consumption.	More investigation is required on transmission reliability between pairs of HCGs and LCGs, especially when link quality deteriorates during subslot S_2 retransmission.
	[15]: IDT-based multi-multicast scheme in multi-cell environment for data rate enhancement using adaptive modulation and coding scheme. Constrained the base stream to satisfy minimum QoS, BER requirement. To avoid excessive transmission failure, scheme uses transmit redundancy to boost LCG data.	One of the few works investigating cooperative multiple transmitter instead of cooperative receivers and exploits multi-cell channel diversity gains.	Requires more efficient method to overcome the huge complexity and synchronization overhead due to different propagation delay of data sent from multiple cells.
	MT and **PF** [21], [26]–[28]: Maximum Throughput and Proportional Fairness. IDT-based. Each user's capacity is a reflection of own channel quality. two-step suboptimal. Adapts information decomposition techniques (MDC and layered coding) to exploit multiuser channel diversity.	Capacity is maximized and fairness is inherently obtained. Ensures fairness in next transmission by remunerating and giving priority to least user in previous transmission.	High coding, decoding complexity and heavy transmission overhead.
MaxMin Fairness	[65] IDT-based, three-Step Suboptimal Algorithm with guaranteed QoS support at the base stream for all users. Provides only subcarrier allocation. Ensure QoS satisfaction for LCG user by adding constraints to the optimization problem.	May result in resource starvation for whole multicast group in low SNR region where LCG user cannot meet QoS requirements.	Potentially more complex than other schemes.
	B-CMS [30]: Integrates PHY layer beamforming and CMS for a two-phase, subgroup transmission. Uses LCGs as benchmark. LCGs use MRC to merge signals from BS and relayed signals from HCGs to achieve higher capacities. Suboptimal due to the relaxation to allow application of gradient guided approximation to obtain solution. Reduces to the classical max-min fairness where minimum utilization is maximized at all iteration.	Use of beamforming and MRC receiver diversity can potentially increase system capacities. Also, CMS reduces computation power demand on the BS.	This scheme shifts intensive computation power demands from the BS and distributes it on the relay nodes. Benefit of this strategy is not obvious considering the limited battery lifespan of mobile relay nodes while BS has higher power supply.

A. Suboptimal Strict Throughput Single-Rate Algorithms

Suboptimal algorithms for multicast-based resource allocation in existing literature can be differentiated based on two basic properties: The reduction assumptions towards simplifying the complexity of the problems, and the isolation methods used to divide the problems into independent steps such that each step has polynomial complexity. For the first, subcarriers are assigned to each group with objective of maximizing total system capacity. This step assumes that total system transmit power is evenly distributed over all subcarriers. In the second, transmit power is optimally allocated to each preselected subcarrier using Lagrange multiplier method or the Karush-Kuhn-Tucker (KKT) conditions [63] - which interestingly, is similar to the conventional waterfilling rule - to enforce group rate proportionality.

These two steps are usually adopted, however, the difference is the techniques used to assign subcarriers, which we shall discuss in detail. For instance, [7], shows that if equal power is applied to selective subcarriers with good channel gains, total throughput of zero pathloss difference of suboptimal heuristic scheme approximates the performance of the optimal scheme even with flat transmit power spectral density (PSD). This simple approach is an example of strict throughput maximization which reduces computation complexity mainly to subcarrier allocation and eliminates need for power allocation. When we consider multiple multicast services, complexity increases because we need to determine which group receives the best subcarrier in each iteration. This subcarrier allocation decision is determined by the system objectives possibly formulated in terms of strict throughput maximization or fairness of access to network resources.

This case was further studied in [19] where subcarrier n is allocated to group g having potential maximum data rate as given in eqn. (8). This is equivalent to assigning subcarrier n to group g with maximum SNR or highest channel gain noting that each group rate is based on worst channel user. Similar works for single multicast MIMO using spatial multiplexing and multi-multicast MIMO using weighting precoding are done in [5] and [16] respectively. These works further define two additional approximations for eqn. (8), which are functions of the eigenvalues of the channel matrix for the low and high SNR regions where the SNR is close to $zero$ and $>>$ average SNR respectively.

$$\max_{g\in G}(R_{g,n}) = |K_g| \log_2\left(1 + \tfrac{1}{N}\hbar_{g,n} P_{Total}\right) \tag{8}$$

The significance of these approximations is the reduction in computation complexity of the algorithm in the SNR extremes. In contrast to LCG user where performance degrades as the number of users increase, these schemes achieve higher system capacities even at the low SNR region since groups with better channel condition always have resources. Results in [5] also show that as the channel power gap increases larger gains is achievable.

As with all STM objective functions, fair access to system resources between groups with diverse carrier-to-noise ratio (CNR) is not considered. If the link difference among multicast group is large, group with high CNR will dominate the resource for a large amount of time, leaving groups with low link quality to starve. For example, schemes in [5] can potentially shut out groups with poor channel gain or fewer numbers of users since it is based on maximum aggregate data rate which increases as users per group and channel gain increase. Tables II shows summary of suboptimal strict throughput single-rate algorithms.

B. Suboptimal Max-Min Fair Single-Rate Algorithms

One possible way to prevent greedy resource utilization by HCG groups and maintain balance between throughput maximization and fairness is to impose minimum number of subcarrier to allocate to each group. Ngo *et al.* [7] shows that by adding one more constraint as shown in eqn. (9) to constraints in (5)-(7), certain level of flexibility and fair resource access can be assured.

$$\sum_{n=1}^{N} \delta_{g,n} \geq \alpha_g|_{g=1}^{G}, \tag{9}$$

where $0 \leq \alpha_g \leq N$, $\sum_{g=1}^{G} \alpha_g \leq N$, and α_g is the minimum number of subcarrier to assign to each group.

Interestingly, the total capacity result of the suboptimal fair scheme approximates performance of the suboptimal strict throughput even at 2.5dB pathloss difference. However, determination of the optimal choice of α_g is not trivial because if $\alpha_g \to 0$, problem becomes strict throughput whereas, the problem becomes strict fairness if $\alpha_g \to \lfloor \frac{M}{G} \rfloor$.

A compensation approach was proposed in [6], [26] where fairness is enforced by compensating each group for low rate - relative to the target rate required by the group - by moving them to better subcarrier with higher CNR. That is, in the next transmission of each multicast group, subcarrier having maximum channel gain (best subcarrier) is assigned to the group with least data rate in the previous transmission. In addition to showing intergroup relationship that may exist in a cell, this approach ensures that no group dominates the system resources and low rate groups do not experience outright resource starvation.

1) Max-Min Fair & QoS Considerations: Besides throughput and fairness system objectives, it is equally important to provide satisfying quality of service to users because bad QoS affects users' level of satisfaction and defeats the system purpose. One way to achieve this is to ensure that achievable rates for each single-rate multicast group satisfy the minimum rate requirements of contents served. In [56], the following constraint is added to constraints (5)-(7):

$$R_{g,n} \geq R_{g,n}^{min} \qquad 1 \leq n \leq N, \tag{10}$$

where $R_{g,n}^{min}$ is the least data rate requirement to satisfy users' QoS requirements. Resulting optimization solution guarantees acceptable service quality if LCG users experience good channel quality but it invariably results in absolute resource starvation for all group members once the minimum user in the group cannot satisfy own rate requirements. Hence, in low SNR regime where LCG user's data rate is less than QoS requirement, system performance degradation may result.

To satisfy the QoS requirements of users in a multicast group, [15], [65] applied IDT and redefined the base stream as the minimum rate all users must receive to satisfy QoS

requirements. Notice that in this case, only the base stream needs to be optimized for QoS as other users can subsequently receive more description to improve service quality. Hence, if a user receives the minimum rate - base stream, QoS is satisfied and additional enhancements simply exploit link condition to provide improved resolutions [36]. Various other constraints can also be added and evaluated for any of the algorithms discussed in this paper.

VI. Multicasting in Current Wireless Standards & Future Directions

Multimedia Broadcast Multicast Services (MBMS):

The Third Generation Partnership Project 2 (3GPP2) Multimedia Broadcast Multicast Services (MBMS) [8], [66] is a unidirectional multicasting services that has been enabled for CDMA2000 [17], [18], 3G Universal Mobile Telecommunications System (UMTS) communication systems and currently been considered for inclusion in the IP-based Mobile WiMAX and LTE. It utilizes point-to-multipoint (PMP) bearer transmission technology, where high-speed multimedia content is delivered from a single source entity to multiple or group of mobile devices or user equipments (UEs). This drastically reduces the linear dependencies between the number of connected UEs and the amount of system resources required.

Besides the addition of new nodes for MBMS implementation across different releases, the major differences between the 3G and 4G MBMS implementations are the enhanced QoS, introduction of link adaptation to improve data rates, reduced communication overhead, and storage requirements. MBMS is envisaged to face stiff competition from competing technologies such as IP-multicast and Digital Video Broadcast - Handheld (DVB-H) which are considered by many to be the closest competitor of MBMS. MBMS provides enhanced security procedures than IP Multicast. Moreover, the *closed* but flexible operational business model of MBMS makes it more appealing to content providers and users unlike the *open* IP multicast where anyone can receive/transmit data sent to a group without authorization and without any form of compensation to the service providers [67]. Nevertheless, there are certain applications for which MBMS is clearly unsuitable whereas IP multicast works seamlessly (e.g. multiuser video conferencing). Similarly, DVB-H provides all the specialized functions of MBMS, however, MBMS can provide uplink services unlike DVB-H which provides only downlink services. Several other technical challenges of MBMS are currently being addressed within the research community.

VII. Potential Research Opportunities

Tremendous research opportunities still exist in multimedia multicasting: First, existing studies in radio resource allocation for multicast broadcast services in OFDMA system have mainly investigated single multicast services, whereas a few, multiple multicast, all solely within a single cell system. This observation raises two issues: Firstly, single multicast and single cell consideration do not adequately capture the essence of multicast systems, efficient multiple multicast or generalized model is still required. Secondly, most existing work on MSRA have focused heavily on subcarrier, power and bit allocation with BER, QoS and transmit power constraints, leaving delay constraints as viable research to explore. Investigating the impact of delay along with other constraints is highly crucial considering the unprecedented amount of delay-sensitive multimedia applications that are envisaged for deployment in the next generation systems.

Thirdly, this study has shown that some works exist on multi-rate cooperative multicasting to reduce power consumption; however, works considering mobile nodes characterized by high resource constraints as described in [58] are still conspicuously lacking. Such evaluation is critical for future communication systems which would predominantly be composed of mobile users. Also related is the evaluation of the impact and complexities of using multiple antenna on mobile terminals for multicasting. Only few authors [5], [25], [68], [69] have considered this challenging task.

Lastly, next generation communication systems such as IEEE802.16m provides massive support for mobile users. Hence, it is not unlikely that multicast group members move rapidly across multiple cells or base stations. Users at the cell edge of one base station may, however, be within good transmission range of another base station transmitting same content and may be able to join another group and share network resource. Similarly, users at boundary may also receive signals affected by co-channel interference from neighboring base stations which further lowers the transmission rate of the received data. Most resource management schemes are developed without consideration of cell interferences and user mobility. Therefore, to enhance cell capacity, more rigorous studies are required on base station cooperation and mobility effect on multicast resource allocation as number of users in groups dynamically changes. Such study may however require cross-layer optimization. Additionally, most MSRA algorithms discussed in this work attempt to offer reliable multicast group transmission which maximize overall system throughput at the expense of complexity, high energy cost, under-utilization of network resources, or trade-off of few selected low gain users. Very efficient solutions are yet to emerge.

Similarly, cooperative multicasting highlighted in this paper still requires significant development. Although, there is a rich literature on physical-layer cooperation that investigate cooperative protocols and relay partner choice selection as shown in [70], however, application to multicast system potentially opens up another research trend. In summary, as demands for multicast services increase, many more practical issues requiring further investigation becomes highly imperative - all offering compelling research topics.

VIII. Conclusion

Multimedia multicasting is a very promising technology for transmission of common multimedia services from a single entity to a group of users using a shared transmission medium. It potentially provides better spectral utilization and has been a focal point of research and development effort over the last several decades. Some of the important problems of multicast systems hindering it from achieving its full potential are, the selection of the most efficient group transmission rates, determination of the optimal MSRA strategy, mechanism

to deal with anomaly behaviors as well as development of techniques to lower associated computational complexities and overheads. In this paper, we have provided a comprehensive survey of these problems and presented a rigorously analyzed multi-dimensional matrix of existing MSRA enabling algorithms proposed in the literature. We also proactively provided a systematic taxonomy of major existing works in multicast resource allocation and presented possible research opportunities to assist interested researchers.

REFERENCES

[1] C. Jie, "Mobile TV - a great opportunity for WiMAX," *Communicate*, no. 41, pp. 34–36, Jun. 2008.

[2] F. Hartung, U. Horn, J. Huschke, M. Kampmann, T. Lohmar, and M. Lundevall, "Delivery of broadcast services in 3G networks," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 188–199, Mar. 2007.

[3] A. M. C. Correia, J. C. M. Silva, N. M. B. Souto, L. A. C. Silva, A. B. Boal, and A. B. Soares, "Multi-resolution broadcast/multicast systems for MBMS," *IEEE Trans. Broadcast.*, vol. 53, no. 1, pp. 224–234, Mar. 2007.

[4] U. Varshney, "Multicast support in mobile commerce applications," *Computer*, vol. 35, no. 2, pp. 115–117, Feb. 2002.

[5] J. Xu, S. Lee, W. Kang, and J. Seo, "Adaptive resource allocation for MIMO-OFDM based wireless multicast systems," *IEEE Trans. Broadcast.*, vol. 56, no. 1, pp. 98–102, Mar. 2010.

[6] K. Bakanoğlu, W. Mingquan, L. Hang, and M. Saurabh, "Adaptive resource allocation in multicast OFDMA systems," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'10)*, Apr. 2010, pp. 1–6.

[7] D. Ngo, C. Tellambura, and H. Nguyen, "Efficient resource allocation for OFDMA multicast systems with fairness consideration," in *Proc. IEEE Radio and Wireless Symposium (RWS'09)*, Jan. 2009, pp. 392–395.

[8] *Introduction of the Multimedia Broadcast/Multicast Service (MBMS) in the Radio Access Network (RAN) Stage 2*, 3GPP TSG TS25. 346, Rev. 6.7.0, Dec. 2005.

[9] T. Jiang, L. Song, and Y. Zhang, *Orthogonal Frequency Division Multiple Access Fundamentals and Applications.* Boston, MA, USA: Auerbach Publications, 2010.

[10] K. Letaief and Y. J. Zhang, "Dynamic multiuser resource allocation and adaptation for wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 38–47, Aug. 2006.

[11] M. Shariat, A. Quddus, S. Ghorashi, and R. Tafazolli, "Scheduling as an important cross-layer operation for emerging broadband wireless systems," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 2, pp. 74–86, Second Quarter 2009.

[12] S. Sadr, A. Anpalagan, and K. Raahemifar, "Radio resource allocation algorithms for the downlink of multiuser OFDM communication systems," *IEEE Commun. Surveys Tuts.*, vol. 11, no. 3, pp. 92–106, Third Quarter 2009.

[13] H. Won, H. Cai, D. Y. Eun, K. Guo, A. Netravali, I. Rhee, and K. Sabnani, "Multicast scheduling in cellular data networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4540–4549, Sep. 2009.

[14] W. Xu, K. Niu, J. Lin, and Z. He, "Resource allocation in multicast OFDM systems: Lower/upper bounds and suboptimal algorithm," *IEEE Commun. Lett.*, vol. 15, no. 7, pp. 722–724, Jul. 2011.

[15] H. Kwon and B. G. Lee, "Cooperative power allocation for broadcast/multicast services in cellular OFDM systems," *IEEE Trans. Commun.*, vol. 57, no. 10, pp. 3092–3102, Oct. 2009.

[16] S. Li, X. Wang, H. Zhang, and Y. Zhao, "Dynamic resource allocation with precoding for OFDMA-based wireless multicast systems," in *Proc. IEEE 73rd VTC Spring*, May 2011, pp. 1–5.

[17] P. Agashe, R. Rezaiifar, and P. Bender, "CDMA2000 high rate broadcast packet data air interface design," *IEEE Commun. Mag.*, vol. 42, no. 2, pp. 83–89, Feb. 2004.

[18] *CDMA2000 High Rate Broadcast-Multicast Packet Data Air Interface Specification*, 3GPP2 3GPP2 C.S0054-0, Rev. 1.0, Feb. 2004.

[19] J. Liu, W. Chen, Z. Cao, and K. Letaief, "Dynamic power and sub-carrier allocation for OFDMA-based wireless multicast systems," in *Proc. IEEE International Conference on Communications (ICC'08)*, May 2008.

[20] C. H. Koh and Y. Y. Kim, "A proportional fair scheduling for multicast services in wireless cellular networks," in *Proc. 64th Vehicular Technology Conference (VTC-'06 Fall)*, Sep. 2006, pp. 1–5.

[21] C. Suh and C. Hwang, "Dynamic subchannel and bit allocation multicast OFDM systems," in *Proc. IEEE 15th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC'04)*, vol. 3, Sep. 2004, pp. 2102–2106.

[22] P. Gopala and H. El Gamal, "On the throughput-delay tradeoff in cellular multicast," in *Proc. International Conf. on Wireless Networks, Comm. and Mobile Computing*, vol. 2, Jun. 2005, pp. 1401–1406.

[23] Y. Sun and K. Liu, "Transmit diversity techniques for multicasting over wireless networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC'04)*, vol. 1, Mar. 2004, pp. 593–598.

[24] A. Narula, M. Lopez, M. Trott, and G. Wornell, "Efficient use of side information in multiple-antenna data transmission over fading channels," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 8, pp. 1423–1436, Oct. 1998.

[25] N. Sidiropoulos, T. Davidson, and Z. Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[26] C. Suh, S. Park, and Y. Cho, "Efficient algorithm for proportional fairness scheduling in multicast OFDM systems," in *Proc. 61st IEEE Vehicular Technology Conference (VTC'05-Spring)*, vol. 3, May 2005, pp. 1880–1884.

[27] C. Suh and J. Mo, "Resource allocation for multicast services in multicarrier wireless communications," in *Proc. IEEE International Conference on Computer Communications (INFOCOM'06)*, Apr. 2006, pp. 1–12.

[28] Changho Suh and Jeonghoon Mo, "Resource allocation for multicast services in multicarrier wireless communications," *IEEE Trans. Wireless Commun.*, vol. 7, no. 1, pp. 27–31, Jan. 2008.

[29] Y. Ma, K. Letaief, Z. Wang, R. Murch, and Z. Wu, "Multiple description coding-based optimal resource allocation for OFDMA multicast service," in *Proc. IEEE (GLOBECOM'10)*, Dec. 2010, pp. 1–5.

[30] T. Han and N. Ansari, "Energy efficient wireless multicasting," *IEEE Commun. Lett.*, vol. 15, no. 6, pp. 620–622, Jun. 2011.

[31] F. Hou, L. Cai, P.-H. Ho, X. Shen, and J. Zhang, "A cooperative multicast scheduling scheme for multimedia services in IEEE 802.16 networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 3, pp. 1508–1519, Mar. 2009.

[32] V. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 74–93, Sep. 2001.

[33] B. Li and J. Liu, "Multirate video multicast over the internet: an overview," *IEEE Network*, vol. 17, no. 1, pp. 24–29, Jan./Feb. 2003.

[34] J. Z. J. Wolf, A. Wyner, "Source coding for multiple descriptions," Bell Labs. Tech. J., pp. 1417–1426, Oct. 1980.

[35] P. Eusebio and A. Correia, "Two QoS regions packet scheduling for multimedia broadcast multicast services," in *Proc. 6th IEE International Conference on 3G and Beyond*, Nov. 2005, pp. 1–5.

[36] S. Deb, S. Jaiswal, and K. Nagaraj, "Real-time video multicast in WiMAX networks," in *Proc. 27th IEEE INFOCOM*, Apr. 2008, pp. 1579–1587.

[37] M. Shao, S. Dumitrescu, and X. Wu, "Layered multicast with inter-layer network coding for multimedia streaming," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 353–365, Apr. 2011.

[38] N. Shrestha, P. Saengudomlert, and Y. Ji, "Dynamic subcarrier allocation with transmit diversity for OFDMA-based wireless multicast transmissions," in *Proc. IEEE Inter. Conf. on Elect. Eng./Elect. Computer Tel. and Information Tech. (ECTI-CON'10)*, May 2010, pp. 410–414.

[39] S. M. Elrabiei and M. H. Habaebi, "Energy efficient cooperative multicasting for MBS WiMAX traffic," in *Proc. IEEE 5th International Symposium on Wireless Pervasive Computing (ISWPC'10)*, May 2010, pp. 600–605.

[40] P. Liu, Z. Tao, Z. Lin, E. Erkip, and S. Panwar, "Cooperative wireless communications: a cross-layer approach," *IEEE Trans. Wireless Commun.*, vol. 13, no. 4, pp. 84–92, Aug. 2006.

[41] Y. Li, G. Su, D. Wu, D. Jin, L. Su, and L. Zeng, "The impact of node selfishness on multicasting in delay tolerant networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2224–2238, Jun. 2011.

[42] C. Long, T. Chen, and X. Guan, "Cooperative amplify and forward in the presence of multiple selfish relays: Performance analysis," in *Proc. Canadian Conf. on Electrical and Computer Engineering (CCECE'08)*, May 2008, pp. 002 045–002 050.

[43] S. Liu, R. Zhang, L. Song, Z. Han, and B. Jiao, "Enforce truth-telling in wireless relay networks for secure communication," in *Proc. IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2011, pp. 1071–1075.

[44] L. Buttyan and J. pierre Hubaux, "Stimulating cooperation in self-organizing mobile ad hoc networks," *ACM/Kluwer Mobile Netw. Appl.*, vol. 8, no. 5, pp. 579–592, Oct. 2003.

[45] S. Zhong, J. Chen, and Y. Yang, "Sprite: a simple, cheat-proof, credit-based system for mobile ad-hoc networks," in *Proc. 22nd IEEE INFOCOM*, 2003, pp. 1987–1997.
[46] P. Michiardi and R. Molva, "A game theoretical approach to evaluate cooperation enforcement mechanisms in mobile ad hoc networks," in *Proc. Modeling Optimiz. Mobile, Ad Hoc,Wireless Netw. (WiOpt)*, 2003.
[47] Wei Yu and K.J.R. Liu, "Game theoretic analysis of cooperation stimulation and security in autonomous mobile ad hoc networks," *IEEE Trans. Mobile Comput.*, vol. 6, no. 5, pp. 507–521, May 2007.
[48] W. Yu and K. Liu, "Secure cooperation in autonomous mobile ad-hoc networks under noise and imperfect monitoring: A game-theoretic approach," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 2, pp. 317–330, Jun. 2008.
[49] W. Lin, H. Zhao, and K. Liu, "Incentive cooperation strategies for peer-to-peer live multimedia streaming social networks," *IEEE Trans. Multimedia*, vol. 11, no. 3, pp. 396–412, Apr. 2009.
[50] Z. Han, Z. Ji, and K. Liu, "A cartel maintenance framework to enforce cooperation in wireless networks with selfish users," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1889–1899, May 2008.
[51] C. Song and Q. Zhang, "Achieving cooperative spectrum sensing in wireless cognitive radio networks," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 13, pp. 14–25, 2009.
[52] B. Niu, H. Zhao, and H. Jiang, "A cooperation stimulation strategy in wireless multicast networks," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2355–2369, May 2011.
[53] I. Toufik and H. Kim, "MIMO-OFDMA opportunistic beamforming with partial channel state information," in *Proc. IEEE International Conference on Communications (ICC'06)*, vol. 12, Jun. 2006, pp. 5389–5394.
[54] B. Wu, J. Shen, and H. Xiang, "Predictive resource allocation for multicast OFDM systems," in *Proc. 5th International Conference on Wireless Communications, Networking and Mobile Computing (WiCom'09)*, Sep. 2009, pp. 1–5.
[55] A. Biagioni, R. Fantacci, D. Marabissi, and D. Tarchi, "Adaptive subcarrier allocation schemes for wireless OFDMA systems in WiMAX networks," *IEEE J. Sel. Areas Commun.*, vol. 27, no. 2, pp. 217–225, Feb. 2009.
[56] W. Xu, Z. He, K. Niu, J. Lin, and W. Wu, "Multicast resource allocation with min-rate requirements in OFDM systems," *Journal of China Universities of Posts and Tel.*, vol. 17, pp. 24–51, 2010.
[57] B. Da and C. C. Ko, "Resource allocation in downlink MIMO-OFDMA with proportional fairness," *Journal of Communications*, vol. 4, pp. 8–13, 2009.
[58] S. Sharangi, R. Krishnamurti, and M. Hefeeda, "Energy-efficient multicasting of scalable video streams over WiMAX networks," *IEEE Trans. Multimedia*, vol. 13, no. 1, pp. 102–115, Feb. 2011.
[59] L. Fortnow, "The status of the P versus NP problem," *Commun. ACM*, vol. 52, pp. 78–86, Sep. 2009.
[60] V. Papoutsis and S. Kotsopoulos, "Chunk-based resource allocation in multicast OFDMA systems with average BER constraint," *IEEE Commun. Lett.*, vol. 15, no. 5, pp. 551–553, May 2011.
[61] V. Papoutsis and S. Kotsopoulos, "Chunk-based resource allocation in distributed MISO-OFDMA systems with fairness guarantee," *IEEE Commun. Lett.*, vol. 15, no. 4, pp. 377–379, Apr. 2011.
[62] V. Corvino, L. Giupponi, A. Perez Neira, V. Tralli, and R. Verdone, "Cross-layer radio resource allocation: The journey so far and the road ahead," in *Proc. 2nd Cross Layer Design, (IWCLD '09)*, Jun. 2009, pp. 1–6.
[63] S. Boyd and L. Vanderberghe, *Convex optimization*, 1st ed. Cambridge University Press, 2004.
[64] M. Li, X. Wang, H. Zhang, and M. Tang, "Resource allocation with subcarrier cooperation in OFDM-based wireless multicast system," in *Proc. IEEE 73rd Vehicular Technology Conference (VTC-'11 Spring)*, May 2011, pp. 1–5.
[65] L. Tian, D. Pang, Y. Yang, J. Shi, G. Fang, and E. Dutkiewicz, "Subcarrier allocation for multicast services in multicarrier wireless systems with QoS guarantees," in *Proc. IEEE Wireless Communications & Networking Conference (WCNC'10)*, Apr. 2010, pp. 1–6.
[66] T. Jiang, W. Xiang, H. Chen, and Q. Ni, "Multicast broadcast services support in OFDMA-Based WiMAX systems [Advances in Mobile Multimedia]," *IEEE Commun. Mag.*, vol. 45, no. 8, pp. 78–86, Aug. 2007.
[67] G. Xylomenos, V. Vogkas, and G. Thanos, "The multimedia broadcast/multicast service," *Wireless Communications and Mobile Computing*, vol. 8, pp. 255–265, 2008.
[68] S. Y. Park and D. J. Love, "Capacity limits of multiple antenna multicasting using antenna subset selection," *IEEE Trans. Signal Process.*, vol. 56, pp. 2524–2534, Jun. 2008.
[69] N. Jindal and Z. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE International Symposium on Information Theory*, Jul. 2006, pp. 1841 –1845.
[70] M. Salem, A. Adinoyi, M. Rahman, H. Yanikomeroglu, D. Falconer, Y.-D. Kim, E. Kim, and Y.-C. Cheong, "An overview of radio resource management in relay-enhanced OFDMA-based networks," *IEEE Commun. Surveys Tuts.*, vol. 2, no. 3, pp. 422–438, Third Quarter 2010.

Richard O. Afolabi received his BSc. degree in Computer Science from The University of Ibadan, Nigeria, in 2004 and MSc. degree in Communications Systems Engineering in 2007 from the Communication and Sensor Networks Lab. (CSNL), School of Information and Communication, Gwangju Institute of Science and Technology (GIST), South Korea. He is currently a research assistant, working towards his Ph.D. degree in the same Lab.

His current research interests include multicasting, radio resource allocation and optimization, cognitive radio and radio access networks in next generation telecommunication systems. He is a graduate student member of the IEEE Communication Society.

Aresh Dadlani earned his BSc. and MSc. degrees in Electrical and Computer Engineering, both from University of Tehran, Iran, in 2007 and 2010, respectively. He was also a research assistant in the School of Computer Science at the Institute for Studies in Theoretical Physics and Mathematics (I.P.M.), Iran from 2008 to 2010. He is currently working towards the Ph.D. degree in Information and Communication at Gwangju Institute of Science and Technology (GIST) in the Republic of Korea.

His research interests include performance modeling and evaluation of complex networks, wireless and mobile communication networks, and Quality of Service (QoS) aspects in WDM optical networks.

Kim Kiseon received the B.Eng. and M.Eng. degrees, in electronics engineering, from Seoul National University, Korea, in 1978 and 1980, and the Ph.D. degree in electrical engineering systems from University of Southern California, Los Angeles, in 1987. From 1988 to 1991, he was with Schlumberger, Houston, Texas. From 1991 to 1994, he was with the Superconducting Super Collider Lab, Texas. He joined Gwangju Institute of Science and Technology (GIST), Korea, in 1994, where he is currently a professor.

His current interests include wideband digital communications system design, sensor network design, analysis and implementation both, at the physical layer and at the resource management layer.