Connection-Level QoS Provisioning in Multiple Transmission Technology-based OFDM System

Youngkyu Choi, Sunghyun Choi
School of Electrical Engineering and INMC
Seoul National University, Seoul, Korea
Email: ykchoi@mwnl.snu.ac.kr, schoi@snu.ac.kr

Sung-Pil Hong
Department of Industrial Engineering
Seoul National University, Seoul, Korea
Email: sphong@snu.ac.kr

***Abstract*— To provide heterogeneous quality-of-service (QoS) to various applications, the bandwidth of orthogonal frequency division multiplexing (OFDM) is divided into multiple number of sub-bands, which employ a different set of transmission technologies, such as multiple-input multiple-output (MIMO), packet scheduling, adaptive modulation and coding (AMC), and hybrid automatic repeat request (HARQ) depending on the traffic type. According to this concept, *DiffSeg* system, where the two-dimensional resource domain of OFDM is filled with four different resource units, called *segments*, and the occupancy ratio of each type of segment is determined by the *segment map*, has been recently proposed in [5][6].**

In this paper, we introduce a concept of *segment diversion* to utilize the radio resource more efficiently considering the status of QoS. We then propose a system optimization model, which finds an optimal segment map and diversion ratios of real-time (RT) connections in order to maximize the residual resource for non-real-time (NRT) traffic while satisfying the minimum QoS requirement of RT traffic. We formulate an optimization problem using mixed integer programming, and then develop two computationally efficient algorithms: simplex-based heuristic and *Maximum Diversion Rule*. Especially, Maximum Diversion Rule is shown to achieve a near-optimal solution with dramatically less complexity.

I. Introduction

An ultimate vision of emerging wireless data communication systems is to enable the users to enjoy what they want anytime anywhere, if possible, with a single device. Accordingly, data communication should be broadband comparably with the wireline Internet, while the traditional voice communication is still well-supported. Reviewing the technologies proposed up to now, real-time (RT) traffic, such as voice, is most likely to match diversity-centric transmission technologies. Since many diversity techniques such as power control, frequency hopping try to compensate the variation of wireless channel, they contribute to satisfying delay constraint of RT applications. Guaranteed resource allocation like circuit switching is preferred to maintain the quality of RT applications.

On the other hand, for non-real-time (NRT) communication such as typical best-effort traffic, opportunistic scheduling, which attempts to send the data at the peak point of channel variation, recently attracted lots of attention. Multi-user diversity resulted from opportunistic scheduling exploits the channel fluctuation rather than mitigates it. Especially, in the packet-based network, sharing the medium brings statistical multiplexing gain, and hence increases the system capacity. Summarizing the above discussion, we can see that the set of transmission technology and resource allocation, which are advantageous to either RT or NRT, is different. This makes it challenging to serve both RT and NRT applications simultaneously in a wireless communication system.

There have been many efforts to support RT applications while pursuing an efficient utilization of wireless medium. The authors in [7] propose to support voice communication by using static priority queue in IEEE 802.11 wireless LAN (WLAN) where distributed coordinated function (DCF) was originally designed without considering RT support. The idea in [8] is to reserve a part of slots during a frame exclusively for RT traffic in EV-DO system. In [9], exponential rule is shown to guarantee the delay requirement of RT traffic statistically while maintaining the high throughput for NRT. In these schemes, the service quality for RT traffic is differentiated or guaranteed through the resource allocation of medium access control (MAC).

Recently, orthogonal frequency division multiplexing (OFDM) and multiple input multiple output (MIMO) technology introduce a new concept of wireless system. OFDM provides multiple channels available at a given time, and MIMO also gives multiple spatial channels, which can be exploited either to increase the transmission rate or to improve the reliability of the transmission depending on its usage. A proper combination of transmission technologies provides an opportunity to build QoS-friendly physical layer (PHY). For example, an 802.16e-based emerging broadband wireless network in Korea, called WiBro [2], divides the entire OFDM bands into two different types of subchannels, i.e., adaptive modulation and coding (AMC) subchannels and diversity subchannels. AMC subchannels aim to use the medium opportunistically to achieve higher spectral efficiency while diversity subchannels are for delivering information more reliably. Another example is differentiated segment-based (DiffSeg) system [5][6]. DiffSeg system defines four different resource units, called *segments*. Each segment uses a different set of transmission technologies, which have relatively superior attributes when sending the traffic with specific QoS requirements[1] to a certain

This research was in part supported by Samsung Electronics and grant No. R01-2005-000-10271-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

[1] According to whether the traffic is sensitive for the delay, the type of traffic is simply classified into two types, i.e., RT or NRT.

user under specific radio environment.[2] The occupancy ratio of each segment type in the resource domain is determined by the *segment map*.

In this paper, we consider both (1) how to select one segment map out of a finite set of pre-determined maps, and (2) for a selected segment map, which segments to use for each RT connection. To the best knowledge of the authors, this problem has not been fully resolved in the OFDM system. As the previous work [9][10][11] has pointed out, the packet scheduling raises several complex issues such as delay or throughput constraint depending on the traffic type. Accordingly, we divide the whole resource allocation problem into two levels; connection-level and packet-level problem. Here, we only deal with the connection-level problem, and leave the packet-level problem as a future research.

The rest of this paper is organized as follows: Section II is devoted to a description of the DiffSeg system which is faithful to the philosophy of multiple transmission technology-based OFDM system. In Section III, we formulate the connection-level resource allocation problem using mixed integer programming that maximizes the residual resource for NRT traffic while satisfying the minimum QoS requirement of RT connections. In Section IV, we propose two computationally efficient algorithms: a simplex method-based algorithm and a practical heuristic algorithm, called *Maximum Diversion Rule*. In Section V, the performance of Maximum Diversion Rule is numerically presented compared with the direct solution of mixed integer programming. Finally, Section VI presents some concluding remarks and future work.

II. DiffSeg System

Basically, the system resource in OFDM[3] consists of two dimensional elements of time and frequency. A certain system such as IEEE 802.11a [1] has one dimensional resource domain, i.e., time, even if OFDM is adopted as its transmission scheme. We exclude such OFDM systems from our interests. Instead, we describe the DiffSeg system briefly in order to clarify the concept of multiple transmission technology-based OFDM system.

In DiffSeg system, the basic unit of transmission is called a *segment*, which comprises a few OFDM symbols and subcarriers. Since multiple types of segments exist, and each segment type is selectively used for intended purposes, the system is named differentiated segment-based (DiffSeg) system. Fig. 1 shows the definition of differentiated segments and a set of associated transmission technologies which the system adopts for each type of segment. Basically, two major criteria are used to classify the segment types: QoS (i.e., RT vs. NRT) and the average carrier to interference ratio (CIR) (i.e., cell center vs. cell boundary.[4]) In terms of QoS, it is the main concern if a connection imposes any latency requirement, and depending on the condition, the connection is classified into either RT or NRT. On the other hand, the average CIR is a representative factor describing the radio environment of the user. It can be interpreted as the distance from the base station (BS), i.e., cell-center or cell-boundary. Besides, the user velocity, inter-cell interference, and antenna correlation are also considered in this criterion. Accordingly, we obtain four different types of segment, and the name of each segment is coined from the combination of letters representing the related criteria.

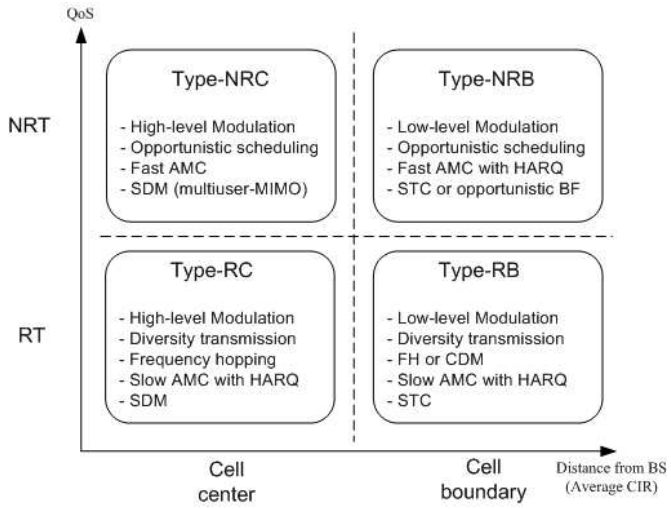


Fig. 1. Definition of differentiated segments. [5]

The reason why each segment type have to use such a set of transmission technologies is described in detail in [5] and [6], and here we just follow their model. However, for the sake of completeness we briefly discuss how to adapt the wireless link for each segment. Basically, NRT segments including type-NRC and type-NRB involve fast AMC while RT segments including type-RC and type-RB employ slow AMC. Fast AMC means that transmission rate is adjusted adaptively to the instantaneous channel quality. Therefore, all the mobile stations (MSTAs) which want to receive the traffic through NRT segments should feedback the instantaneous channel quality via uplink signaling. Consequently, the BS can schedule the transmission of segment toward increasing the spectral efficiency. Additionally, multiple spatial channels provided by multiuser-MIMO (see type-NRC segment in Fig. 1) give more freedom to use the resource in an opportunistic manner. On the other hand, slow AMC means that rate adaptation occurs infrequently in order to cope with the change of average CIR, which is expected to reflect the effects from both pathloss and shadowing. On the other hand, the fast fading due to multi-path is mitigated via diversity techniques such as frequency hopping and MIMO, and then the reliability of transmission is achieved using hybrid automatic repeat request (HARQ) with incremental redundancy.

[2]The factors characterizing users' radio environment include the signal to interference ratio (SINR), Doppler frequency, antenna coherence, and so on. For detailed rationales, refer to [5] and [6].

[3]In this paper, we focus on the downlink. However, the study can be extended similarly to uplink, i.e., orthogonal frequency division multiple access (OFDMA).

[4]The CIR value fluctuates over time irrespective of a mobile's geographical location within the cell. Accordingly, this location-based grouping is only for an illustrative purpose.

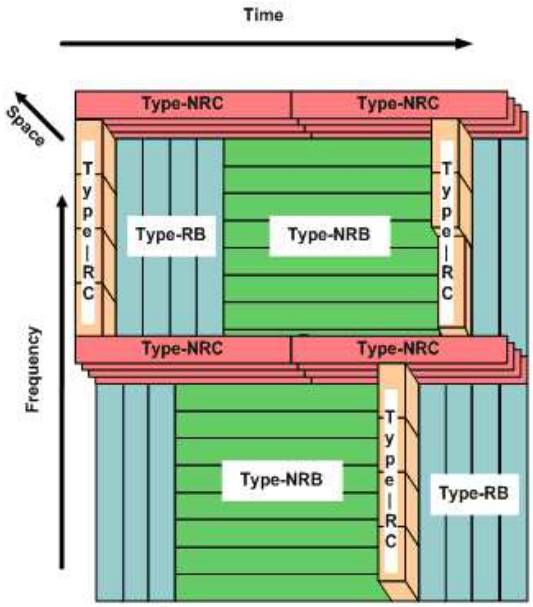


Fig. 2. An exemplified arrangement of segments at a specific segment map. [5]

Another important point is that each segment type has different shape of rectangle whose width and height corresponds to the number of OFDM symbols and subcarriers, respectively. These different shapes result from the consideration of several trade-offs: frequency diversity, the loop delay of HARQ, signaling burden due to fast AMC, and so on. Specifically, RT segments occupy smaller number of OFDM symbols and larger amount of subcarriers. In Fig. 2, the length of RT segments in time is $1/8$ times that of NRT segments. Note that both RT and NRT segments have the same amount of resource, i.e., the area of rectangle, nevertheless.

The whole resource domain can be covered by a specific mixture of segments, where a *segment map* uniquely determines the physical mapping of segments to subcarrier and OFDM symbol. Fig. 2 shows an example where four different types of segments are disposed for a given segment map. For type-NRC and RC segments, we see that there are a stack of rectangles along the space axis, which represents the spatial channels generated by MIMO technology. When we define a time interval between two consecutive broadcast intervals as a MAC frame, the pattern of segments is repeated every MAC frame until a new segment map is chosen. Practically, the total number of segment maps is finite, and the *map index*, which identifies a segment map, is known to all MSTAs within a cell through broadcasting. Therefore, when a segment is allocated to an MSTA, and this is notified through control signaling, the MSTA is ready to receive it using appropriate transceiver technologies corresponding to the type of a segment in Fig. 1.

III. Connection-level Resource Allocation

In this section, we propose a model that maximizes resource allocation for NRT connections while guaranteeing minimum QoS to the RT connections. In doing so, we introduce a concept of *segment diversion* for an efficient resource utilization. The system model is then formulated as a mathematical optimization problem.

TABLE I
Grouping of real-time connections

RT connection sets	Corresponding MSTA's on-going connections
Λ_1	RC
Λ_2	RC, NRC
Λ_3	RC,NRB
Λ_4	RB
Λ_5	RB,NRC
Λ_6	RB,NRB

A. System Model

We denote a set of all the available segment maps by Φ. Then, a specific segment map with index k, ϕ_k $(\in \Phi)$ is uniquely described by $\phi_k = [\phi_{1k}, \phi_{2k}, \phi_{3k}, \phi_{4k}]^T$ $(k = 1, 2, \ldots, |\Phi|)$, where ϕ_{jk} represents the occupancy ratio of the j-th type segment $(j = 1, 2, 3, 4)$ in the k-th segment map. We denote RC, RB, NRC, and NRB by indexes 1, 2, 3 and 4, respectively. The incoming connections are classified into RT or NRT depending on the QoS requirement of the associated application. This is analogous to the functionality of the convergence sublayer in IEEE 802.16 specification [3] in that the characteristic of an incoming connection is interpreted into an internal connection type defined by a specific wireless interface. By default, RT and NRT packets are transmitted via the corresponding type of segments, separately. Let us call this type of segment allocation the *static* method.

We now introduce an alternative segment allocation, called *segment diversion*, in which RT packets are (partly) diverted to NRT segments. Compared with the former static allocation, the segment diversion provides an additional degree of freedom in resource allocation. Since it is possible to share divided medium, i.e., RT and NRT segments, we have a chance to exploit the radio resources more efficiently. Ideally, if there are infinitely many segment maps, then a continuous adaptation of a segment map would achieve the optimal resource allocation, and segment diversion might need not be considered. The segment diversion enables a finite set of pr-determined segment maps to handle the variable demands of RT and NRT traffic more efficiently.

Here, we should note that every RT connection can not be diverted. Since NRT segments require instantaneous channel feedback for opportunistic scheduling as noted in Section II, RT connections of an MST, which has no concurrent NRT connection, can not be diverted. Therefore, an RT connection should be differently dealt with depending on whether the corresponding MSTA has any NRT connection simultaneously. Since whether the segment type for RT and NRT traffic is classified into cell center or cell boundary, i.e., type-(N)RC or type-(N)RB is determined by independent criteria such as CIR, Doppler frequency, and antenna coherence, an MSTA can have both RC connections and NRB connections concurrently, for example. Accordingly, an RT connection can be classified into one of six disjoint sets Λ_i (for $i = 1, \ldots, 6$) in Table I. For instance, any MSTA with RT connections belonging to Λ_2

also has at least an NRT connection, which is classified into type-NRC segment, simultaneously. The complete set of RT connections in a cell is represented by $\Lambda = \bigcup_{i=1}^{6} \Lambda_i$.

Further, we assume that the BS has complete information about the QoS status which is provided to all the RT connections by the currently available RT and NRT segments. The current QoS reward vector of the i-th RT connection is represented by

$$r_i = \begin{bmatrix} r_{i1} \\ r_{i2} \end{bmatrix},$$

where r_{i1} and r_{i2} are the QoS rewards expected when all the packets of the connection are transmitted exclusively via RT and NRT type segments, respectively. In other words, r_{i2} represents the QoS status of the RT connection when the packets of the connection are fully diverted to NRT type segments. Naturally, all RT connections belonging to set Λ_1 and Λ_4 have $r_{i2} = 0$ since they can not be diverted. Note that MSTAs are expected to update r_i via uplink signaling in a periodic or aperiodic manner. How the QoS of RT applications could be quantified is investigated in [14] and the references therein, and hence the QoS rewards could be kept track of. Simply, one of the criteria determining the QoS status of an RT application can be the latency. Given the function of user-perceived QoS vs. average latency for an RT application, r_i can be quantified by measuring the latency. Since the QoS reward of an RT connection is typically a non-increasing function of the latency, r_{i1} is likely larger than r_{i2} achieved by an opportunistic scheduling, which is usually inferior in terms of latency than other fair scheduling such as round-robin algorithm [9]. However, many different versions of opportunistic scheduling can compromise on the latency performance. Therefore, packet scheduling for NRT-type segments can be another factor affecting r_{i2}. However, we do not deal with this issue here since we focus on the connection-level QoS provisioning in this paper.

Similarly, the transmission rate vector of the i-th RT connection is represented by

$$\alpha_i = \begin{bmatrix} \alpha_{i1} \\ \alpha_{i2} \end{bmatrix},$$

where α_{i1} is the transmission rate supported via an RT segment and α_{i2} is the *average* rate achieved when scheduled in an opportunistic manner via the associated NRT segments. Here, the unit of a transmission rate is defined as the number of bits deliverable using a single segment ($bits/segment$). Since the opportunistic scheduling can achieve a higher data rate in an average sense, α_{i2} is typically larger than α_{i1}.

In the connection-level QoS provisioning, our objective is to find the optimal index of segment map and diversion ratio for each RT connection such that they maximize the residual bandwidth for NRT traffic while satisfying the minimum QoS requirement of RT traffic. Fig. 3 illustrates the block diagram of the system structure where the segment map decision block implements the algorithm solving our optimization problem. The diversion ratios determined by the segment map decision block are used as parameters by the packet scheduler. The map

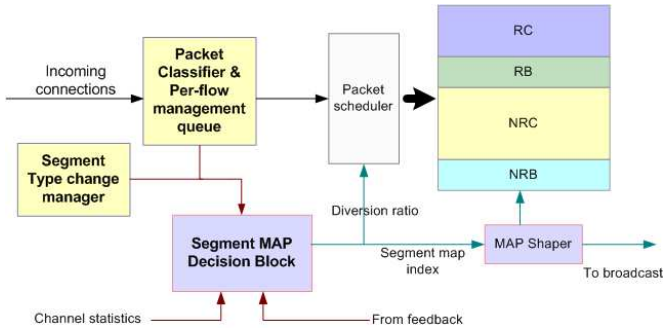


Fig. 3. Structured system block.

shaper in the PHY changes the map configuration according to the segment map index obtained from the segment map decision block. Then, the newly-determined map index is broadcast to all MSTAs to synchronize the downlink resource map. *Packet classifier* block is in charge of mapping an incoming connection to either type of RT or NRT. The *segment type change manager* block keeps track of the desirable segment types for RT and NRT traffic of each MSTA, respectively. Apparently, the overall system structure includes a cross-layered attribute, and the map decision block plays a key role by determining the system efficiency.

B. The optimization formulation

As discussed in Section III-A, given the set Λ of RT connections, the objective is to find an optimal map ϕ_k from Φ and each RT connection's diversion ratio that maximize the residual resource for the NRT connections. Let b_i $(bits/frame)$ be the bandwidth requirement of the i-th RT connection and $n_i = [n_{i1}, n_{i2}]^T$ be a vector of ratios by which the packets from the i-th RT connection are conveyed through RT and NRT segments via diversion, respectively. Also, let X be the total number of available segments within a MAC frame, and it is constant for any segment map.

Then, our problem is to find

$$R := \max\{R_k : \phi_k \in \Phi\}, \tag{1}$$

where R_k is the optimum value of the k-th subproblem defined by Eqs. (2)–(9).

$$R_k := \max\left\{X(\phi_{3k} + \phi_{4k}) - \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} \frac{b_i}{\alpha_{i2}} n_{i2}\right\} \tag{2}$$

s.t

$$\textstyle\sum_{i\in\Lambda_2\cup\Lambda_3} \frac{b_i}{\alpha_{i1}} n_{i1} \le X\phi_{1k} - \sum_{i\in\Lambda_1} \frac{b_i}{\alpha_{i1}}, \tag{3}$$

$$\textstyle\sum_{i\in\Lambda_5\cup\Lambda_6} \frac{b_i}{\alpha_{i1}} n_{i1} \le X\phi_{2k} - \sum_{i\in\Lambda_4} \frac{b_i}{\alpha_{i1}}, \tag{4}$$

$$\textstyle\sum_{i\in\Lambda_2\cup\Lambda_5} \frac{b_i}{\alpha_{i2}} n_{i2} \le X\phi_{3k}, \tag{5}$$

$$\textstyle\sum_{i\in\Lambda_3\cup\Lambda_6} \frac{b_i}{\alpha_{i2}} n_{i2} \le X\phi_{4k}, \tag{6}$$

$$r_{i1}n_{i1} + r_{i2}n_{i2} \ge \rho_i,\ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6, \tag{7}$$

$$n_{i1} + n_{i2} = 1,\ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6, \tag{8}$$

$$n_{i1} \ge 0,\ \ n_{i2} \ge 0,\ \ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6. \tag{9}$$

In Eq. (1), we choose, from the pre-determined set of segment maps, the map with the maximum residual bandwidth. For each map ϕ_k, the objective function in Eq. (2) is the residual

resource for NRT traffic, i.e., the sum of two NRT-type segments minus the amount occupied by diverted RT connections. Therefore, the objective for the k-th subproblem is to find the vectors n_i for $\forall i \in \Lambda_2 \cup \Lambda_3 \cup \Lambda_5 \cup \Lambda_6$ which maximize the residual resources under the map ϕ_k. Eqs. (3) and (4) enforce the RT bandwidth constraints while Eqs. (5) and (6) enforce the NRT bandwidth constraints. Finally, Eq. (7) means when a RT connection is (partially) diverted to NRT segments, the aggregated QoS should be guaranteed in a normalized sense. That is, the linear combination of the QoS rewards attained via RT and NRT segments, respectively, should not be less than the minimum QoS requirement, ρ_i.

Apparently, under the static allocation, i.e., $n_{i1} = 1, n_{i2} = 0$, for every $i \in \Lambda$, the optimization problem reduces to a simple problem of finding map ϕ_o with the minimum sum of ϕ_{1o} and ϕ_{2o} that satisfies

$$X\phi_{1k} \geq \textstyle\sum_{i\in\Lambda_1\cup\Lambda_2\cup\Lambda_3} \frac{b_i}{\alpha_{i1}}, \text{ and} \tag{10}$$

$$X\phi_{2k} \geq \textstyle\sum_{i\in\Lambda_4\cup\Lambda_5\cup\Lambda_6} \frac{b_i}{\alpha_{i1}}. \tag{11}$$

Note that there may be multiple optimal maps ϕ_o, which have the same values of both ϕ_{1o} and ϕ_{2o} since ϕ_{3o} and ϕ_{4o} of ϕ_o are not decided. (This can actually happen in the segment diversion case as well.) In such a case, we can apply an additional criterion to the ratios of NRT segments, i.e., ϕ_{3o} and ϕ_{4o}.

A possible policy is to allocate the resource proportionally to the ratio of input to output rates, where the input rate of an NRT segment type is the aggregated rate of the NRT connections with the segment type, and the output rate is the average transmission rate supported by the segment type, respectively. Here, the average transmission rate is affected by both packet scheduling policy and rate adaptation algorithm. However, since the packet-level problem is not explicitly dealt with in this paper, some weights μ and $1-\mu$ are assumed given to type-3 and type-4 segments, respectively. Therefore, the residual segments are allocated proportionally to the weight, and hence the optimal map index can be uniquely determined. At this time, we should take care of the minimum amount of resource occupied by diverted RT connections as shown by both Eqs. (5) and (6). Accordingly, the tentative values of ϕ_{3o} and ϕ_{4o} can be decided as follows:

$$\widetilde{\phi}_{3o} = \mu\tfrac{R}{X} + \phi_{3o,low}, \tag{12}$$

$$\widetilde{\phi}_{4o} = (1-\mu)\tfrac{R}{X} + \phi_{4o,low}, \tag{13}$$

where $\phi_{3o,low}$ and $\phi_{4o,low}$ are the lower bounds given by Eqs. (5) and (6). Finally, we can choose a map ϕ_o with ϕ_{3o} and ϕ_{4o} nearest to $\widetilde{\phi}_{3o}$ and $\widetilde{\phi}_{4o}$.

Back to the segment diversion problem, let w_{i1} and w_{i2} denote the bandwidth of the i-th RT connection split into RT and NRT segments, respectively:

$$w_{i1} := b_i/\alpha_{i1},\ w_{i2} := b_i/\alpha_{i2}\ (\text{segments/frame}). \tag{14}$$

Using Eqs. (8) and (14), we can simplify the k-th subproblem as follows.

$$\begin{aligned} &R_k + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} w_{i2} \\ =\ &\max\ X(\phi_{3k}+\phi_{4k}) + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} w_{i2}n_{i1} \end{aligned}$$

s.t

$$\begin{aligned} X\phi_{1k} - \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3} w_{i1}n_{i1} &\geq \textstyle\sum_{i\in\Lambda_1} w_{i1}, \\ X\phi_{2k} - \textstyle\sum_{i\in\Lambda_5\cup\Lambda_6} w_{i1}n_{i1} &\geq \textstyle\sum_{i\in\Lambda_4} w_{i1}, \\ X\phi_{3k} + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_5} w_{i2}n_{i1} &\geq \textstyle\sum_{i\in\Lambda_2\cup\Lambda_5} w_{i2}, \\ X\phi_{4k} + \textstyle\sum_{i\in\Lambda_3\cup\Lambda_6} w_{i2}n_{i1} &\geq \textstyle\sum_{i\in\Lambda_3\cup\Lambda_6} w_{i2}, \\ (\rho_i - r_{i2})/(r_{i1}-r_{i2}) \leq n_{i1} &\leq 1,\ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6. \end{aligned}$$

Finally, renaming

$$\begin{aligned} \gamma_i &:= (\rho_i - r_{i2})/(r_{i1}-r_{i2}), \\ m_i &:= n_{i1} - \gamma_i, \\ \Pi_{jk} &:= X\phi_{jk},\ j = 1,2,3,4, \end{aligned}$$

we get a more compact form, Problem 1.

Problem 1: LP_k**, the k-th subproblem**

$$\begin{aligned} &R_k + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} (1-\gamma_i)w_{i2} \\ =\ &\max\ \Pi_{3k} + \Pi_{4k} + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} w_{i2}m_i \end{aligned}$$

s.t

$$\begin{aligned} \Pi_{1k} - \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3} w_{i1}m_i &\geq \textstyle\sum_{i\in\Lambda_1} w_{i1} + \sum_{i\in\Lambda_2\cup\Lambda_3} \gamma_i w_{i1}, \\ \Pi_{2k} - \textstyle\sum_{i\in\Lambda_5\cup\Lambda_6} w_{i1}m_i &\geq \textstyle\sum_{i\in\Lambda_4} w_{i1} + \sum_{i\in\Lambda_5\cup\Lambda_6} \gamma_i w_{i1}, \\ \Pi_{3k} + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_5} w_{i2}m_i &\geq \textstyle\sum_{i\in\Lambda_2\cup\Lambda_5} (1-\gamma_i)w_{i2}, \\ \Pi_{4k} + \textstyle\sum_{i\in\Lambda_3\cup\Lambda_6} w_{i2}m_i &\geq \textstyle\sum_{i\in\Lambda_3\cup\Lambda_6} (1-\gamma_i)w_{i2}, \\ 0 \leq m_i &\leq 1-\gamma_i,\ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6. \end{aligned}$$

Denote this linear program by LP_k. The optimum value R of Eq. (1) can be obtained by solving $|\Phi|$ linear program subproblems $\{\mathrm{LP}_k : k \in \Phi\}$, where $|\cdot|$ represents the number of elements in a set. This straightforward method is, however, not desirable for any purpose as the size of Φ may be large. Therefore, we need to consider an alternative formulation of the problem in order to utilize optimization solution techniques.

IV. Computationally Efficient Solution Methods

In this section, we reformulate the mathematical optimization model described in Section III-B into a mixed integer program. Then we discuss how to find an exact optimal solution efficiently by exploiting the structure of the reformulation referred to as MIP. Notice that the optimization model as well as MIP are valid only when the set Λ and the parameters of each connection $i \in \Lambda$ remain unchanged. In principle, the optimality can be maintained in a dynamic case where existing connections terminate, new connections arrive, or the rewards vector r_i change, by optimizing MIP redefined over updated Λ and related parameters over time. However, although MIP is solvable in semi-real time, it is still not an option to rely on the exact solutions of MIP to keep up with the real-time changes of the system due to its computation complexity.

Instead, we propose two heuristics. The first algorithm is essentially a dynamic adaptation of simplex method via the

useful tool of Lagrangian multipliers (or dual variables). In this paper, we just sketch the idea of the algorithm rather than fully elaborate or implement it. The second heuristic is more practical and problem-specific. It is a greedy-type algorithm which, at each step, fully divert RT connections as far as the QoS requirement is not violated. As demonstrated in Section V, it provides a near-optimal solution in real time. In doing so, the heuristic solutions are compared with the exact optima of MIP computed off-line.

A. An Exact Algorithm

By introducing $0-1$ variables $\{y_i \in \{0,1\} : i \in \Phi\}$, we can integrate our problem into a single mixed integer program, Problem 2.

Problem 2: **MIP**

$$
\begin{aligned}
& R + \textstyle\sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6}(1-\gamma_i)w_{i2} = \\
& \max \ \textstyle\sum_{k\in\Phi}(\Pi_{3k}+\Pi_{4k})\,y_k + \sum_{i\in\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6} w_{i2}m_i \\
& \text{s.t} \\
& \textstyle\sum_{k\in\Phi}\Pi_{1k}y_k - \sum_{i\in\Lambda_2\cup\Lambda_3} w_{i1}m_i \geq \\
& \qquad\qquad \textstyle\sum_{i\in\Lambda_1} w_{i1} + \sum_{i\in\Lambda_2\cup\Lambda_3}\gamma_i w_{i1}, \\
& \textstyle\sum_{k\in\Phi}\Pi_{2k}y_k - \sum_{i\in\Lambda_5\cup\Lambda_6} w_{i1}m_i \geq \\
& \qquad\qquad \textstyle\sum_{i\in\Lambda_4} w_{i1} + \sum_{i\in\Lambda_5\cup\Lambda_6}\gamma_i w_{i1}, \\
& \textstyle\sum_{k\in\Phi}\Pi_{3k}y_k + \sum_{i\in\Lambda_2\cup\Lambda_5} w_{i2}m_i \geq \sum_{i\in\Lambda_2\cup\Lambda_5}(1-\gamma_i)w_{i2}, \\
& \textstyle\sum_{k\in\Phi}\Pi_{4k}y_k + \sum_{i\in\Lambda_3\cup\Lambda_6} w_{i2}m_i \geq \sum_{i\in\Lambda_3\cup\Lambda_6}(1-\gamma_i)w_{i2}, \\
& \textstyle\sum_{k\in\Phi} y_k = 1 \\
& y_k \in \{0,1\},\ k\in\Phi, \qquad (15) \\
& 0 \leq m_i \leq 1-\gamma_i,\ i \in \Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6.
\end{aligned}
$$

This might not appear as a progress since a mixed integer program is, in general, much more difficult than linear program. However, note that MIP has a special structure:

1) MIP has only 5 constraints. Hence, any optimal basic feasible solution of the linear program obtained by relaxing the integrality constraint (15), will have at most 5 non-zero variables. Considering other continuous variables m_i's competing for basic variables, it is most likely that only two or three y_k's will have non-zero variables. This means the linear program relaxation LP gives a very tight and practical approximation of MIP.
2) MIP has a lot more variables, namely, $|\Phi|$ $0-1$ variables plus $|\Lambda_2\cup\Lambda_3\cup\Lambda_5\cup\Lambda_6|$ continuous variables than constraints. It is well-known that in practice the computation time of simplex algorithm is proportionate to the number of constraints [12]. Furthermore, we may accelerate the computation by combining the revised simplex method and the column generation instead of considering the whole set of columns explicitly in each iteration.

From these, one can see that linear program relaxation not only provides a tight and practical approximation of MIP solution, but also offers a basis of an efficient exact algorithm. In the light of this, a standard integer programming solution method relying on linear programming relaxation, such as Branch-and-Bound method, is likely to produce an optimal solution a lot faster than the naive method of solving all the $|\Phi|$ linear subproblems, $\mathrm{LP_k}$. Indeed, we could confirm this from computational experiment. A straightforward implementation of Branch-and-Bound method delivers an optimal solution in a few number of branchings, and hence in time of solving a few linear programs.

B. Simplex-based Heuristic

When there is a change, such as a connection arrival, a connection termination, or a change in the reward vector in an on-going operation, it corresponds to a relatively small portion of system parameters. More specifically, it corresponds to insertion or deletion of a column (a variable) and/or a small change in the right-hand-side of the constraints of the mathematical model, MIP in Problem 2. For instance, if new connection i belongs to Λ_1, there is no other choice than to transmit it via RC segment. Thus, we can see from MIP or more conveniently from Fig. 4, an illustration of the coefficient structure, that only the right-hand-side of the first constraint is incremented by w_{i1}. Intuitively, the new optimum will not be far from the old one: Most values of y_k and m_i's will remain the same in the new optimum solution. In the light of this observation, it is desirable to devise an algorithm which updates the solution with a minimal computation while compromises the optimality minimally.

Once the system is in operation with the $\bar{k}$-th segment map, namely $y_{\bar{k}} = 1$, one can regard MIP as a linear program with the variables m_i ($i \in \Lambda$). It is well-known that the *Lagrangian multiplier* or the *dual variable* of the current basic feasible solution (not necessarily optimal) are used to calculate the reduced cost of each non-basic variable in the improving iterations of the simplex method [13]. The reduced cost is the increasing rate of the objective value when the variable increases. For each multiplier of the five constraints is the change rate of the objective value per unit change in the right-hand-side of the corresponding constraint. This set of multipliers, a 5-dimensional vector z, can be uniquely determined from the working basis of the current solution. Now we sketch a heuristic adaptation of this simplex method principle which maintains the system optimality in the dynamic situation. In doing so, for simplicity, we will assume that due to the admission control the resource is enough for the existing and incoming connections.

1) Changes in connections: Suppose the i-th connection has terminated. Then, maintaining the current solution for the rest of the connections is most likely the best strategy as a small increment of the residual resource will not modify the optimal solution significantly. (There needs to be a technical consideration to keep a full-dimensional basis if m_i was a basic variable. But, a minor step can resolve this and will not be discussed in detail.)

Suppose, on the other hand, a new connection i is incoming. If i belongs Λ_1 or Λ_4, as discussed above, the only choice is to transmit it via RC or RB segments, respectively. It will results in a small increment in the right-hand-side (see Fig.

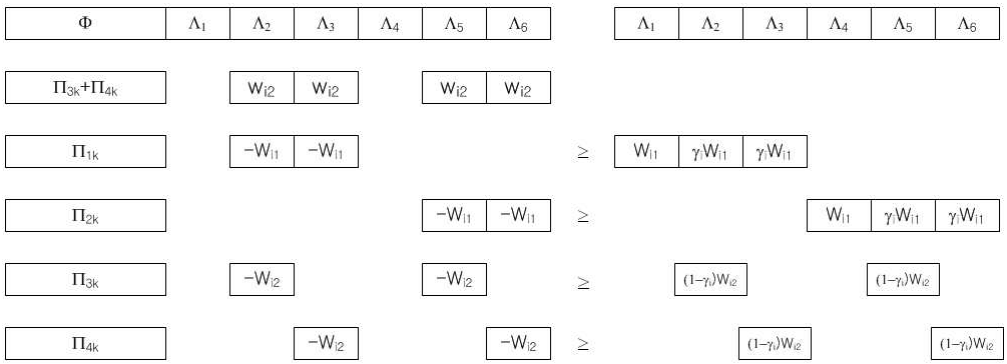


Fig. 4. Coefficient structure of MIP

4) and again retaining the current solution of the on-going connections will not affect the optimality significantly. In other cases, namely when $i \in \Lambda_j$ with $j = 2, 3, 5, 6$, the new connection will also create a new column in the left-hand-side of the constraints. Then, we can calculate the reduced cost for m_i as in the simplex method using the new column and the Lagrangian multipliers z. By adopting the well-known simplex method rule, we can set $m_i = 0$ if the reduced cost is non-positive, and $m_i = 1 - \gamma_i$ if non-negative.

2) Changes in rewards: When the reward vector of an existing connection changes, it will modify the corresponding columns as discussed in Section IV-B.1. Therefore, one can apply a rule which is similar to the one we adopted for a new connection.

3) Updating map ϕ: In the above discussion, we assumed that the segment map is fixed to some $\phi_{\bar{k}}$. But, if the suboptimality of the current map is a significant level, we may choose a different map from Φ, which improves the objective value: For each map ϕ_k, we write

$$\triangle_k := \Pi_k - \Pi_{\bar{k}}. \tag{16}$$

Then define

$$\delta_k = \triangle_{3k} + \triangle_{4k} - \sum_{j=1}^{4} \triangle_{jk} z_j. \tag{17}$$

Consider an imaginary variable with the constraint coefficient column $\triangle_k$ and the objective coefficient $\triangle_{3k} + \triangle_{4k}$. Then, it is easy to see that if the variable has value 0 (1), then it corresponds to the current map $\phi_{\bar{k}}$ (the new map ϕ_k). δ_k is actually the reduced cost of the imaginary variable of map ϕ_k. Therefore, (17) is the change rate of the objective value when the map changes from $\phi_{\bar{k}}$ to ϕ_k. Naturally, we can choose k with the largest value of δ_k. We note, however, that it is an over-estimation, as the objective value is a piecewise linear concave function of the change in the right-hand-side, and hence the change rate holds only over some interval. One possible solution for this problem is to compensate the coefficients with more accurate values by solving the exact algorithm periodically.

C. Max Diversion Rule

Now we propose a greedy-type heuristic for MIP, which attempts to fully divert RT connections as long as the QoS constraints are satisfied. To this end, we use the following proposition.

Proposition 1: The maximum net gain achievable from a segment diversion of the i-th RT connection is given as

$$\mathcal{D}(i) = b_i \left(\frac{1}{\alpha_{i1}} - \frac{1}{\alpha_{i2}} \right) \left(\frac{r_{i1} - \rho_{min,i}}{r_{i1} - r_{i2}} \right). \tag{18}$$

Proof: Given n_{i2}, the diversion will save $\frac{b_i}{\alpha_{i1}} n_{i2}$ from the resource amount occupied by RT-type segments. Instead, $\frac{b_i}{\alpha_{i2}} n_{i2}$ of NRT-type segments is needed due to RT connection's diversion. Therefore, the net saving of resource is $b_i(\frac{1}{\alpha_{i1}} - \frac{1}{\alpha_{i2}})n_{i2}$. Since α_{i2} is likely larger than α_{i1}, the net saving is hopefully positive. Also, from Eqs. (7) and (8), the maximum of n_{i2} is $\frac{r_{i1} - \rho_{min,i}}{r_{i1} - r_{i2}}$. ■

To justify an RT connection's diversion, the data rate of NRT-type segment should be greater than that of RT-type segment. If so, Proposition 1 implies that the net benefit from diversion is proportional to RT traffic's bandwidth b_i and the maximum of n_{i2}, which is a function of parameters related with QoS, i.e., r_i and $\rho_{min,i}$. The basic idea of this heuristic is diverting all the RT connections of which $\mathcal{D}(i)$ is positive.

Algorithm 3: **Max Diversion Rule**

Step 0: Compute an optimal segment map ϕ^s for the static allocation.

Step 1: For each $i \in \Lambda_2 \cup \Lambda_3 \cup \Lambda_5 \cup \Lambda_6$, compute the value $\mathcal{D}(i)$ from Eq. (18). If $\mathcal{D}(i)$ is positive, set $n_i = \left[\frac{\rho_{min,i} - r_{i2}}{r_{i1} - r_{i2}}, \frac{r_{i1} - \rho_{min,i}}{r_{i1} - r_{i2}} \right]^T$. Otherwise, set $n_i = [1, 0]^T$.

Step 2: From the current value of n_i for $i \in \Lambda$, find a map ϕ_d from Φ with the RT segment ratios, ϕ_{1d} and ϕ_{2d}, that maximizes the objective function of Eq. (2).

Step 3: If $\phi_{1d} = \phi_{1s}$ and $\phi_{2d} = \phi_{2s}$, then set $n_i = [1, 0]^T$, $\forall i \in \Lambda$ and go to Step 5.

Step 4: Otherwise, sort the RT connections in the intersection of $\{i | \mathcal{D}(i) > 0\}$ and $i \in \Lambda_2 \cup \Lambda_3 \cup \Lambda_5 \cup \Lambda_6$ in a non-increasing order of $\mathcal{D}(i)$'s, and refer to the

sorted set as Θ.[5] For $j = 1, \cdots, |\Theta|$, define $\Theta_j \subset \Theta$ to be the set of the first j elements of Θ. Find the minimum j that results in the same RT ratios as ϕ_{1d} and ϕ_{2d}. Finally, $\forall i \in (\Lambda_2 \cup \Lambda_3 \cup \Lambda_5 \cup \Lambda_6) \setminus \Theta_j$, set $n_i = [1, 0]^T$.

Step 5: Using Eqs. (12) and (13), determine a unique segment map in Φ.

As the initial map, the algorithm uses ϕ^s, an optimal map attained by the static allocation (Step 0). Given the initial map, it finds all the RT connections whose $\mathcal{D}$ is positive and fully diverts them to NRT-type segments (Step 1). Then, algorithm adjust the map: It finds the best map ϕ^d under the maximum diversion ratios obtained in Step 1 (Step 2). If the diversion does not contribute to increasing the NRT-type resource, i.e.,$\phi^s = \phi^d$, it would be better not to apply a segment diversion since it makes the RT connection's QoS compromised without returns. Therefore, if the map coincides the initial map, then terminate the procedure (Step 3). Otherwise, the diversion ratios achieving ϕ^d are found based on a rule of the largest $\mathcal{D}$ first: The RT connections are sorted in a non-increasing order of $\mathcal{D}$. Then, fully divert the first j connections until the total portion of RT-type segments reaches $\phi_{1d} + \phi_{2d}$. Similar to Step 3, we should avoid excessive segment diversion not to compromise the RT connection's QoS without returns if the target map is reached. (Step 4). Finally, considering parameter μ, we can obtain the tentative values for ϕ_{3o} and ϕ_{4o}, respectively. As discussed in Section III-B, a unique map, ϕ_o, can be determined using these values (Step 5).

In Step 0 and Step 2, the algorithm searches through Φ to find an appropriate map. This can be done in $\mathcal{O}(|\Phi|)$ time. Another major operation is to sort the set Θ in Step 4. The maximum cardinality of Θ is $|\Lambda_2 \cup \Lambda_3 \cup \Lambda_5 \cup \Lambda_6|$. Hence, the proposed algorithm has the complexity of $\mathcal{O}(\max[|\Phi|, |\Theta| \log |\Theta|])$. Therefore, the running time of the heuristic is likely to be very short. Indeed, as shown in Section V, the Max Diversion Rule provides a near-optimal segment map and connection diversion ratios with extremely reduced complexity.

V. Computational Experiment

A. Segment MAP Configuration

Our discussion so far is based on four different segment types. However, Fig. 1 shows that RC and RB type segments employ almost the same set of transmission technologies. In fact, the essential difference is the number of antennas: An MSTA with multiple antennas uses the spatial division multiplexing (SDM) while an MSTA with single antenna adopts the space time coding (STC). Therefore, in practice, we can handle RC and RB types in the same way. This reduces the dimension of segment MAP space from four to three. This also implies that there can be only three sets of RT connections in Table I, i.e., $\Lambda_1 - \Lambda_3$.

[5]If there exist multiple connections with the same value of $\mathcal{D}(i)$, break the ties by a random experiment for the connections to be diverted with an equal probability.

TABLE II
Modulation & Coding Scheme (MCS)

MCS index	Modulation	Code rate	Size ($bits/segment$)
1	QPSK	3/8	336
2	QPSK	3/4	672
3	16QAM	9/16	1008
4	16QAM	3/4	1344
5	64QAM	5/8	1680
6	64QAM	3/4	2016

More specifically, we consider a pre-determined configuration of segment map as follows. A MAC frame, spanning between two broadcast intervals, is covered by 23 sub-patterns called *cluster*, and a cluster is a collection of 184 segments. Hence, deciding to choose which cluster is equivalent to optimizing the entire map. A cluster can be configured using the following rule: First, NRT and RT-type segments are provisioned with one of 24 possible ratios, i.e., $\xi : 23 - \xi$ ($\xi = 0, 1, 2, \cdots, 23$), respectively. Then, the NRT-type segments are provisioned again for NRC-type and NRB-type segment with one of 9 possible ratios, i.e., $\psi : 8 - \psi$ ($\psi = 0, 1, 2, \cdots, 8$). According to this rule, we can obtain 208 different maps in total, counting one for $\xi = 0$. Given map index i ($i = 1, 2, \cdots, 208$), the associated parameters, $\xi(i)$ and $\psi(i)$ can be written as follows:

$$\xi(i) = \begin{cases} 0 & , i = 1, \\ \lfloor \frac{i-2}{9} \rfloor + 1 & , i \neq 1, \end{cases} \tag{19}$$

$$\psi(i) = (i-2) - 9\lfloor \tfrac{i-2}{9} \rfloor, \quad i \neq 1. \tag{20}$$

In summary, $23 \cdot \xi(i)\psi(i)$ corresponds to NRC-type segments, $23 \cdot \xi(i)(8 - \psi(i))$ does to NRB-type segments, and $23 \cdot (184 - 8\xi(i))$ does to RT-type segments, respectively. Accordingly, Π_{ik} in Problem 2 are replaced by the above expressions, and it can be rewritten as Problem 4.

Problem 4: **MIP with specific map configuration**

$$R + \textstyle\sum_{i \in \Lambda_2 \cup \Lambda_3} (1 - \gamma_i) w_{i2} =$$
$$\max \ \textstyle\sum_{k=1}^{208} 23 \cdot 8\xi(k) y_k + \sum_{i \in \Lambda_2 \cup \Lambda_3} w_{i2} m_i$$

s.t

$$\textstyle\sum_{k=1}^{208} 23 \cdot (184 - 8\xi(k)) y_k - \sum_{i \in \Lambda_2 \cup \Lambda_3} w_{i1} m_i \geq \sum_{i \in \Lambda_1} w_{i1} + \sum_{i \in \Lambda_2 \cup \Lambda_3} \gamma_i w_{i1},$$

$$\textstyle\sum_{k=1}^{208} 23 \cdot \xi(k)\psi(k) y_k + \sum_{i \in \Lambda_2} w_{i2} m_i \geq \sum_{i \in \Lambda_2} (1 - \gamma_i) w_{i2},$$

$$\textstyle\sum_{k=1}^{208} 23 \cdot \xi(k)(8 - \psi(k)) y_k + \sum_{i \in \Lambda_3} w_{i2} m_i \geq \sum_{i \in \Lambda_3} (1 - \gamma_i) w_{i2},$$

$$\textstyle\sum_{k=1}^{208} y_k = 1, \ y_k \in \{0, 1\}, \ k \in \Phi,$$
$$0 \leq m_i \leq 1 - \gamma_i, \ i \in \Lambda_2 \cup \Lambda_3.$$

B. Numerical Results

Table II shows the modulation and coding schemes (MCS's) used in the experiment. Specifically, the MCS's are designed so that a segment can deliver integer multiples of the basic payload amount, namely, 336 bits. For RT traffic, we consider video streaming, which is encoded with MPEG4 part10-AVC, and requires the average bandwidth of 512 $kbits/s$.

TABLE III

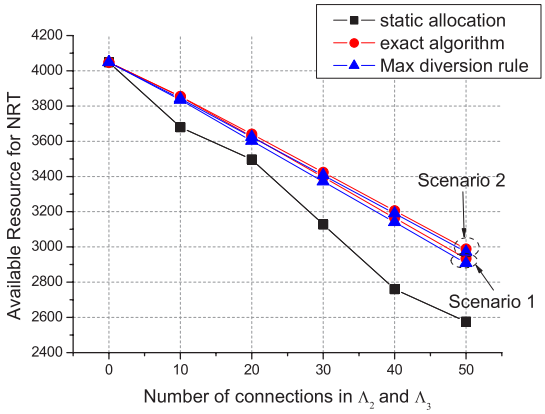
PARAMETERS FOR CONNECTION-LEVEL RESOURCE ALLOCATION

Set of RT connections	$[\alpha_{i1}, \alpha_{i2}]$	$[r_{i1}, r_{i2}]$
Λ_1	336, N/A	1, N/A
scenario 1 - Λ_2	672, 1344	1, 0.8
- Λ_3	336, 1008	1, 0.7
scenario 2 - Λ_2	672, 2016	1, 0.8
- Λ_3	336, 1344	1, 0.7
scenario 3 - Λ_2	672, 1344	1, 0.5
- Λ_3	336, 1008	1, 0.3

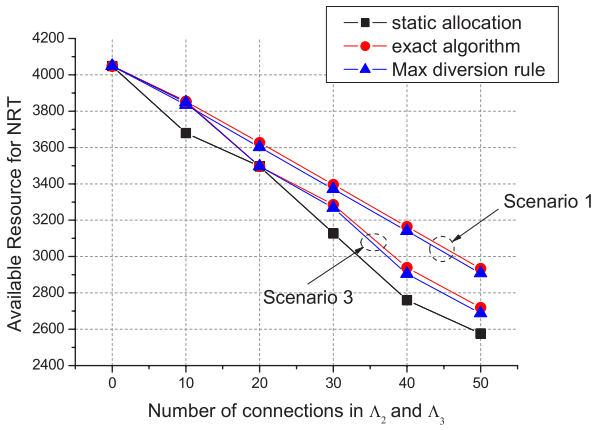
When a MAC frame has time duration of 13.15 ms, the video traffic should be transmitted with the rate of 6,733 $bits/MACframe$. The minimum QoS requirement for video is set to $\rho_{min,video} = 0.9$. We evaluate the performance of each allocation policy using three scenarios as shown in Table III. For each scenario, we assume that all RT connections belonging to a given set Λ_i have the same characteristics in terms of QoS reward r_i and transmission rates α_i as described in Table III. Additionally, we further assume that the QoS of RT traffic can be perfectly supported by RT-type segments, that is, $r_{i1} = 1$ when RT traffic is transmitted using only RT-type segments. Across all the scenarios, the characteristics for the connections in Λ_1 are fixed.

Fig. 5 shows the amount of residual resources available for NRT traffic for a various number of RT connections. For a given value on X axis, both $|\Lambda_2|$ and $|\Lambda_3|$ are equally set to the value while $|\Lambda_1|$ is fixed to five. Note that the maximum objective value is equal to 4,232($= 23\,\text{clusters} \times 184\,\text{segments/cluster}$) when the whole resources of a MAC frame are devoted to NRT-type segments. We first observe that the segment diversion provides more resources to NRT traffic than the static allocation for every scenario. We also find that Max Diversion Rule provides a fairly good solution compared with that from the exact MIP solution. Even at the worst case, the solution from Max Diversion Rule does not deviate from 1% range of MIP's. Fig. 5(a) compares the results of scenarios 1 and 2. As shown in Table III, scenario 2 has higher data rates of NRT type segments over scenario 1, but scenario 2 brings just a marginal increase of residual resources compared with scenario 1. Fig. 5(b) presents the effects from current QoS rewards, r_i. As the QoS rewards supported by NRT-type segments is degraded at scenario 3, there is little room for segment diversion. Therefore, the residual resource approaches to the result of the static allocation. The reason why the resource gap between scenarios 1 and 3 is larger than that between scenarios 1 and 2 can be understood as follows. Recalling Eq. (18), the diversion effect depends on both the data rate and current QoS rewards. When we define a relative diversion effect between two scenarios i and j as $\frac{\mathcal{D}_{scn\,i}}{\mathcal{D}_{scn\,j}}$, a simple calculation shows that $\frac{\mathcal{D}_{scn\,1}}{\mathcal{D}_{scn\,3}} \simeq 2.43$ is larger than $\frac{\mathcal{D}_{scn\,2}}{\mathcal{D}_{scn\,1}} \simeq 1.18$.

Next, we consider a different situation where the characteristics of RT connections, namely, b_i and $\rho_{min,i}$, are different. For this purpose, we additionally consider a popular RT application, i.e., voice over IP (VoIP). According to the G. 711, we assume that VoIP traffic requires the average bandwidth of 64



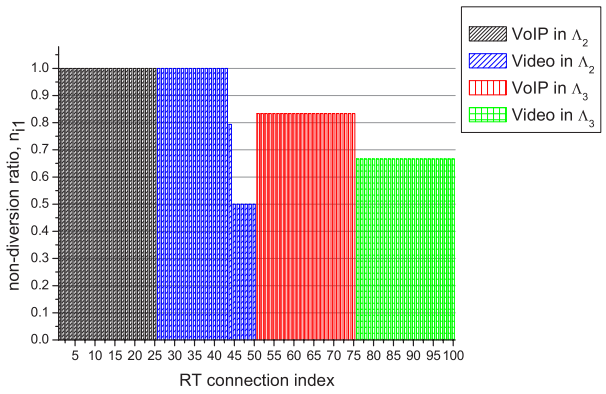
(a) Comparison of scenario 1 and scenario 2, i.e., the effect from data rate.



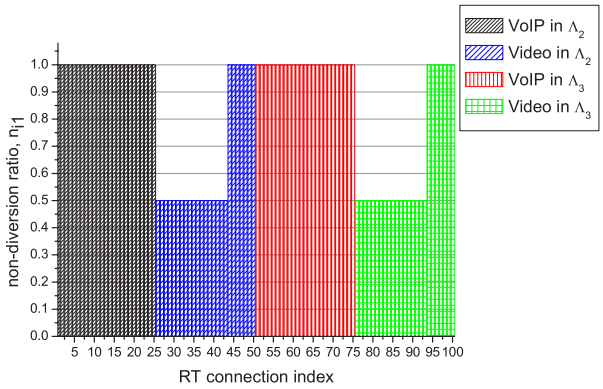
(b) Comparison of scenario 1 and scenario 3, i.e., the effect from QoS status.

Fig. 5. Available resources for NRT vs. Number of RT connections when $|\Lambda_1| = 5$.

$kbits/s$, which is equivalent to 842 $bits/frame$, and demands a higher minimum QoS requirement than video streaming, i.e., $\rho_{min,voip} = 0.95$. Basically, the parameters of scenario 1 in Table III are used. For total 100 RT connections, the same number of RT connections are used for VoIP and video streaming, respectively: The connections with indexes $1 \sim 25$ are VoIP connections belonging to set Λ_2, and those with indexes $26 \sim 50$ are video streaming connections belonging to set Λ_2, too. For the half of RT connections, the RT connections with indexes $51 \sim 75$ of Λ_3 are VoIP connections, and the rest with indexes $76 \sim 100$ are video streaming connections. In this configuration, the static allocation, the exact algorithm, and Max Diversion Rule yield the objective values of 3128, 3420, and 3395, respectively. The exact algorithm and Max Diversion Rule choose the same ξ value, which means that the same segment map is chosen. However, as shown in Fig. 6(a) and (b), non-diversion ratios, i.e., n_{i1} for individual connections, are determined differently. As addressed in Section IV-C, Max Diversion Rule diverts a RT connection with higher $\mathcal{D}(i)$ first,



(a) non-diversion ratio, n_{i1}, when using the exact algorithm.



(b) non-diversion ratio, n_{i1}, when using Max Diversion Rule.

Fig. 6. Comparison of non-diversion ratio between the exact algorithm and Max Diversion Rule.

and stops diversion as soon as reaching the target segment map. Due to this reason, only the video streaming connections with higher $\mathcal{D}(i)$ are diverted as shown in Fig. 6(b). Recall that $\mathcal{D}(i)$ is proportional to the incoming rate, i.e., b_i. On the other hand, the exact algorithm tries to find all the candidates for diversion in order to minimize the empty room of RT-type segments as possible. Accordingly, parts of the VoIP connections belonging to Λ_3 are also diverted.

Based on this observation, we may devise an advanced version of Maximum Diversion Rule. Right before the target map is accomplished, we may try to find other candidates for diversion, which minimize the empty room of RT-type segments. It can be easily implemented by slightly changing the rule of the largest $\mathcal{D}(i)$ first. However, we expect that the improvement will be marginal since Maximum Diversion Rule already achieves the near-optimal performance.

Lastly, we measure the computation time on a computer with the clock speed of 2.0 GHz. For 105 RT connections, the exact algorithm takes more than 200 ms while Max Diversion Rule requires very short time less than 1 μs. The absolute time values do not mean much since they heavily depend on the underlying computation platform. However, the reduction of the computation time by 1/200,000 implies that it will be easier to run Max Diversion Rule at a run-time. Note that in the real system, the segment map and diversion ratios should be adapted over time in per frame basis.

VI. Conclusion

In this paper, we address how to maximize the resource for NRT traffic while satisfying the minimum QoS requirement of RT connections in a multiple transmission technology-based OFDM system such as DiffSeg. For this purpose, we devise a concept of segment diversion, and formulate an optimization problem based on MIP. Since the exact solution for MIP is computationally intensive while the algorithm should run at a run-time of the system operation, we propose two heuristic algorithms: simplex-based algorithm and Max Diversion Rule. Especially, through numerical results, Max Diversion Rule is shown to achieve a near-optimal solution with a remarkably reduced computation complexity. Finally, our future work is to tackle the second phase work of resource allocation, i.e., packet-level QoS provisioning.

References

[1] IEEE Std 802.11a, *Supplement to Part 11: Wireless LAN Medium Access Control (MAC) and Phisical Layer (PHY) specifications: High-speed Phisical layer in the 5 GHz Band*, IEEE Std 802.11a-1999, 1999.

[2] TTAS.KO-06.0082, *Specifications for 2.3GHz band Portable Internet Service - Physical & Medium Access Control Layer*, Standard, TTA, June 2005.

[3] IEEE 802.16, *IEEE Standard to Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems*, October 2004.

[4] IEEE 802.16e/D12-2005, *Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems: Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands*, October 2005.

[5] June Moon, Jae-Yun Ko, and Yong-Hwan Lee, "Differentiated Segments based OFDM System for the Next-Generation Radio Access System," in *Proc. of IST Mobile & Communications Summit*, June 19–23, 2005.

[6] June Moon, Jae-Yun Ko, and Yong-Hwan Lee, "An Air Interface Framework Design for the Next-Generation Radio Access System," to appear in *IEEE Journal on Selected Areas in Communications*, 2nd quarter, 2006.

[7] Jeonggyun Yu, Sunghyun Choi, and Jaehwan Lee, "Enhancement of VoIP over IEEE 802.11 WLAN via Dual Queue Strategy," in *Proc. of IEEE ICC'04*, Paris, France, June 2004.

[8] Youngjune Choi and Saewoong Bahk, "Scheduling for VoIP Service in cdma2000 1x EV-DO," in *Proc. of IEEE ICC'04*, June 2004.

[9] S. Shakkottai and A. Stolyar, "Scheduling Algorithms for a Mixture of Real-Time and Non-Real-Time Data in HDR," in *Proc. of the 17th International Teletraffic Congress (ITC-17)*, September 2001.

[10] Matthew Andrews, Krishnan Kumaran, Kavita Ramanan, Alexander Stolyar, Phil Whiting, and Rajiv Vijayakumar, "Providing Quality of Service over a Shared Wireless Link," *IEEE Communications Magazine*, vol. 39, pp. 150–154, February 2001.

[11] Xin Liu, Edwin K. P. Chong, and Ness B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 2053–2064, October 2001.

[12] Karl Heinz Borgwardt, "Some distribution-independent results about the asymptotic order of the average number of pivot steps of the simplex method," *Mathematics of Operations Research*, vol. 7, no. 3, pp. 441–462, 1982.

[13] Katta G. Murty, *Linear Programming*, Wiley, 1983.

[14] A. Kajackas, V. Batkauskas, and A. Medeisis, "Individual QoS Rating for Voice Services in Cellular Networks," *IEEE Communications Magazine*, vol. 42, no. 6, pp. 88–93, June 2004.