Practical Scheduling Algorithms with Contiguous Resource Allocation for Next-Generation Wireless Systems

Shu Sun and Sungho Moon

Abstract—This paper proposes three novel resource and user scheduling algorithms with contiguous frequency-domain resource allocation (FDRA) for wireless communications systems. The first proposed algorithm jointly schedules users and resources selected adaptively from both ends of the bandwidth part (BWP), while the second and third ones apply disjoint user and resource selection with either single-end or dual-end BWP strategies. Distinct from existing contiguous FDRA approaches, the proposed schemes comply with standards specifications for fifth-generation (5G) and beyond 5G communications, and have lower computational complexity hence are more practical. Simulation results show that all of the proposed algorithms can achieve near-optimal performance in terms of throughput and packet loss rate for low to moderate traffic load, and the first one can still perform relatively well even with a large number of users.

Index Terms—Beyond 5G (B5G), quality of service (QoS), frequency-domain resource allocation (FDRA), scheduling.

I. INTRODUCTION

In a wireless communications system with a next-generation NodeB (gNB) and multiple user equipments (UEs), a pivotal issue to tackle is the scheduling of available time and frequency resources to the UEs in order to satisfy certain quality of service (QoS) requirements such as throughput, fairness, latency, and/or reliability. According to the specifications by the 3rd Generation Partnership Project (3GPP) [1], [2], there are two types of downlink frequency-domain resource allocation (FDRA): type 0 and type 1. In downlink FDRA of type 0, the resource block (RB) assignment information comprises a bitmap indicating the resource block groups (RBGs) that are allocated to the scheduled UE, where an RBG consists of a set of consecutive virtual RBs defined by higher layer parameters [1]. In downlink FDRA of type 1, the RB assignment information signifies to a scheduled UE a set of contiguous allocated non-interleaved or interleaved virtual RBs within the active bandwidth part (BWP) [1]. Two key discrepancies between type-0 and type-1 FDRA are (1) type 0 is on the RBG level while type 1 is on the RB level, and (2) the resources (RBGs or RBs) assigned to each UE can be non-contiguous for type 0, while they must be contiguous for type 1.

A variety of scheduling methods for contiguous FDRA have been proposed previously [3]–[8], predominantly for single-carrier frequency division multiple access in the Long-Term-Evolution-Advanced system. An optimal algorithm was presented in [3], which yielded the best performance but was quite intricate. To reduce the complexity, a greedy heuristic allocation was then proposed in [3] which performed adjacent RB allocation expansion around a localized optimal RB for each UE. At each iteration, the UE and feasible RB combination arousing the largest increase in weighted capacity was selected. In [4], a two-step FDRA scheme was proposed prioritizing the most demanding UEs in terms of their QoS requirements. The authors of [5] proposed a suboptimal algorithmic solution to address the problem of ergodic sum-rate maximization with the constraint of consecutive RB allocation, where the performance gap to optimal solution was limited to 10%. The invention in [6] also contained two steps where the allocation leading to maximum throughput was first found without considering the contiguity constraint, which was then iteratively refined to reach an allocation satisfying contiguity. The algorithm presented in [7] was based on channel gain matrix and iterative RB cluster selection with highest channel gain. More prior art can be found in [8]. The existing strategies, however, mainly concentrate on throughput and do not take into account other QoS criteria such as delay and packet drop rate [3]–[7], and/or involve high computational complexity when the number of UEs is large [3], [6].

In this article, we put forth three scheduling algorithms with type-1 (i.e., contiguous) FDRA which are aligned with 3GPP standards for fifth-generation (5G) and beyond-5G (B5G) communications [1], [2] and have relatively low computational complexity thus viable to deploy in practice. In the first proposed algorithm, FDRA and UE scheduling are jointly conducted to achieve near-optimal performance, whereas in the second and third proposed algorithms, UE selection is executed first, followed by RB allocation, whose major advantage is low complexity. Moreover, all of the proposed algorithms are flexible in terms of scheduling metric such as sum-rate, proportional fairness (PF) [9] and modified largest weighted delay first (M-LWDF) [10]. System-level simulations are carried out to validate and compare the performance of the proposed algorithms, using several traffic types with diverse packet sizes, arrival rates, and QoS requirements.

The authors are with the Next Generation and Standards Group, Intel Corporation, Santa Clara, CA 95054 USA (e-mail: shu.sun@intel.com; sungho.moon@intel.com).

This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.
II. SYSTEM MODEL AND PROBLEM FORMULATION

In this work, we investigate a downlink cellular system comprising of one gNB and K UEs indexed by the set \( K = \{0, \ldots, K-1\} \), where the UEs’ traffic types can be heterogeneous with dissimilar QoS requirements. The transmission BWP \( W \) is orthogonally divided into \( B \) RBs indexed by the set \( B = \{0, \ldots, B-1\} \). The payload for UE \( k \) is denoted by \( L_k \). There are two constraints in type-1 FDRA in 3GPP 5G and B5G specifications \([1], [2]\): (1) exclusivity, meaning an RB can only be allocated to at most one UE; (2) contiguity, i.e., the RBs assigned to each UE must be contiguous. Fig. 1 illustrates the input and output relation per slot in the UE and resource scheduling process with type-1 FDRA \([1], [2]\). The input incorporates the UE set \( K \), RB set \( B \), payload and channel state information (CSI) per UE, where the CSI usually embodies rank indicator (RI), precoding matrix indicator (PMI), and channel quality indicator (CQI) \([1]\). The output includes the selected UE set \( K^* \), selected RB set \( B^* \) implied by resource indication value (RIV) per selected UE, and RI, PMI, and modulation and coding scheme (MCS) per selected UE, where an RIV corresponds to a starting virtual RB and a length pertaining to contiguously allocated RBs \([1]\).

In some of the proposed algorithms to be elaborated in Section III, the calculation of transport block size (TBS) \([1]\) over a certain number of RBs is needed. If wideband (WB) CSI \([1]\) is available, where WB denotes the entire active BWP, the TBS is directly computed using the WB CSI. If subband (SB) CSI is available (while RI is still WB) \([1]\), where an SB is equivalent to an RBG herein, the TBS over all the selected RBs (or RBGs) is calculated via the procedure below: (1) Convert each SB CQI to SB MCS, (2) compute the effective MCS over all the selected RBs (or RBGs), and (3) calculate the TBS over all the selected RBs (or RBGs) using the effective MCS and WB RI. In this article, the effective MCS equals the median of the MCSs over all the selected RBs (or RBGs), but it can also be the average, maximum, or other quantities related to SB MCS. Further, to obtain WB CQI to be utilized in some of the proposed algorithms, the effective MCS is computed over the entire active BWP, which is then converted back to CQI.

For UE \( k \) on RB \( b \), given the estimated channel \( H_{k,b} \) and precoding matrix codebook \([1]\), the RI, PMI, and MCS can be obtained via Algorithms 1 and 2 in \([1]\), after which TBS\(_{k,b}\) is calculated via the method mentioned above. The achievable rate of UE \( k \) on RB \( b \) in each slot is \( r_{k,b} = \text{TBS}_{k,b} \). Let \( B_k \) denote the set of RBs allocated to UE \( k \), the achievable rate of UE \( k \) is \( r_k = \sum_{b \in B_k} r_{k,b} \). The scheduling metric (e.g., sum-rate, PF, M-LWDF) can be flexible depending on the system requirement. Considering QoS requirement, we select M-LWDF as the scheduling metric as an example, which is expressed as \([10]\) \( \mu_{k,b} = -\left(\log \delta_k/\tau_k\right)d_k r_{k,b}/R_k \), where \( \delta_k \), \( \tau_k \), \( d_k \), and \( R_k \) denote the acceptable packet drop probability, delay threshold (the maximum allowable delay from packet generation to packet scheduling), head-of-line (HOL) delay, and historical average rate of UE \( k \), respectively. The optimization problem is formulated as

\[
\begin{align*}
\text{max} & \quad \left\{ \sum_{k \in K} \sum_{b \in B} \mu_{k,b} \right\}_{0 \leq A} \subseteq A \\
\text{subject to} & \quad B_k \cap B_{k'} = 0, \forall k \neq k', k' \in K, \\
& \quad d_k \leq \tau_k, \forall k \in K,
\end{align*}
\]

where \( A \) is the set of all possible RB allocations satisfying the contiguity constraint. (P1) is non-convex whose optimal solution requires exhaustive search with prohibitively high computational complexity. Therefore, in the next section, we propose three practical sub-optimal algorithms to tackle (P1).

III. SCHEDULING ALGORITHMS

A. Proposed Three Scheduling Algorithms

Since it is almost impossible to obtain the optimal solution to (P1) with reasonable computational complexity, we propose three sub-optimal algorithms to solve (P1). Inspired by the observation that \( \sum_{k \in K} \sum_{b \in B_k} \mu_{k,b} \) in \([1]\) is likely to be maximized if the UEs who yield the largest \( \sum_{b \in B_k} \mu_{k,b} \), while consuming the minimum resources are scheduled first, we propose an algorithm named Joint Allocation with Dual Ends (JADE), whose procedures are detailed in Algorithm 1. Essentially, JADE jointly prioritizes the UE and RB(s) in each allocation step that produces the largest scheduling metric with the minimum number of RBs, where the RB selection is performed and compared between both ends of the active BWP to take advantage of frequency diversity. It is worth noting that a variant of JADE, where the RBs are allocated from only one end, rather than both ends, of the BWP, can be applied as well. As the variant is likely to yield inferior performance to JADE due to less frequency diversity, its performance is not shown herein.

Note that in JADE, the number of TBS and scheduling metric calculation is proportional to \( K^2 \) due to the iteration for each remaining UE and RB. To further reduce the computational complexity, two lower-complexity algorithms are designed, i.e., Disjoint Allocation with Single End (DASE) and Disjoint Allocation with Two Ends (DATE). The main design principle of DASE and DATE is to guarantee the QoS for UEs with the most stringent delay and acceptable packet drop probability requirements. In both DASE and DATE, UE selection is done first based on their delay and acceptable packet drop probability requirements as well as the number of RBs needed, followed by RB selection. The only difference between DASE and DATE lies in that RB selection is conducted from only one end of the BWP in DASE, while both ends of the BWP are considered and compared for RB selection in DATE. Detailed steps for DASE and DATE are provided in Algorithm 2 and Algorithm 3, respectively.

B. Scheduling Algorithm with Type-0 FDRA

Ideally, the performance of the proposed algorithms should be compared with that of the optimal type-1 FDRA which,
However, requires exhaustive search over UEs, the starting position of RBs per UE, and the number of RBs per UE, whose complexity is prohibitively high and hence almost impossible to realize. On the other hand, although also requiring exhaustive search, the optimal type-0 FDRA is possible to realize capitalizing on a different and smaller search space. Additionally, if the same frequency granularity is assumed for both type-0 and type-1 FDRA, the optimal type-0 allocation is expected to be superior to the optimal type 1, since discontinuous FDRA enjoys higher flexibility in terms of best UE and resource combination selection. To this end, we compare the performance of the proposed scheduling algorithms with one using the optimal type-0 FDRA which serves as a benchmark. In type-0 FDRA, the scheduling metric is first calculated per UE per RBG, then the UE and RBG combination corresponding to the largest scheduling metric is selected for scheduling and excluded from further selection afterwards in the current slot. Subsequently, the scheduling metric is recalculated per UE per RBG, followed by best UE and RBG combination selection and exclusion, so on and so forth, until there is no remaining UE or RBG.

C. Complexity Analysis

Besides the RBG-level type-0 FDRA, we also compare the proposed algorithms with a representative sub-optimal contiguous FDRA algorithm in the industry published in [3], which is named Localized Expansion of Adjacent Positions (LEAP) herein, to evaluate the performance enhancement by the proposed algorithms over LEAP. The weighted capacity metric in LEAP is replaced by the scheduling metric for fair comparison. Assuming each UE needs \( M \) RBs on average, and the number of RBs in an RBG is \( M_{\text{RB}} \), the computational complexity of all the considered algorithms is provided in Table I.

As expected, type-0 FDRA possesses the highest complexity due to exhaustive search over all UE and RBG combinations. In a typical example where \( K = 30, B = 270, M = 10, \) and \( M_{\text{RB}} = 4 \), the complexity of LEAP is slightly lower than that of FDRA, both of which are approximately an order of magnitude lower than type-0 FDRA. DATE has the lowest complexity which is slightly lower than that of DASE, since these two algorithms exploit disjoint UE and RB selection.

IV. Simulation Results

System-level simulations are conducted to assess and compare the performance of the proposed three type-1 FDRA algorithms. Table I lists the simulation settings, where the traffic models comprise both eMBB (enhanced Mobile Broadband) and URLLC (Ultra-Reliable Low-Latency Communications) (including arVr2, powerDist2, and ITS [15], where arVr2 denotes the second type of augmented reality/virtual reality, powerDist2 represents the second type of power distribution grid fault and outage management, and ITS stands
for intelligent transportation system \cite{13}). The total number of UEs in our simulations vary from 10 to 50, and the ratios of eMBB, arVr2, powerDist2, and ITS UEs are about 1:1:1:1. Table III details the parameters for the traffic models used in our simulations. We note that URLLC traffic can also be scheduled by puncturing the ongoing eMBB transmission, but that scheme has its own drawbacks as well and is out of the scope of this paper whose overarching focus is the scheduling of different traffic types using the same time resource.

The overall and two close-up views of packet throughput for various traffic types are illustrated in Fig. 2 and Fig. 3 respectively. The maximum throughput degradation of the proposed algorithms against the type-0 algorithm and maximum throughput gain over LEAP are summarized in Table IV. Fig. 4 shows the packet loss rate, where a packet is considered lost if it is not entirely scheduled before reaching its delay threshold. For LEAP, the swift increase in packet loss at 30 UEs in Fig. 4 for ITS and arVr2 unveils its instability and sensitivity to the UE amount and/or locations. The following key observations can be drawn from these results:

1) In general, JADE outperforms DASE, DATE as well as LEAP. The superiority of JADE over DASE and DATE is especially evident when the number of UEs is large (e.g., 50), as shown by the throughput and packet loss rate for 50 UEs in Fig. 2 to Fig. 4. Comparing JADE and LEAP, as shown in Fig. 3 and Fig. 4, JADE yields higher throughput and lower packet loss rate in most cases, with a maximum throughput gain of 9.9% (see Table IV). The reason is that LEAP allocates RBs locally around the first best RB for each UE, but if the channel quality happens to change abruptly around the first RB, the overall channel quality may degrade hence incurring performance loss.

2) JADE has comparable and sometimes even better performance in contrast with the type-0 algorithm, as shown by the throughput for arVr2 UEs in Fig. 3 since type-0 FDRA is based on the RBG level (4 RBs per RBG in the simulation) while type 1 is of RB-level which has a finer frequency granularity hence higher flexibility. The maximum throughput degradation of JADE against type-0 FDRA is only 0.9%.

3) Dual-end FDRA surpasses its single-end counterpart, as validated by DATE and DASE, due to its higher frequency diversity.

4) As demonstrated by the left plot of Fig. 3, Table IV, and Fig. 4, DASE and DATE perform well and even exceed LEAP when the number of UEs is not very large, i.e., up to 40 UEs in this case. Since these two algorithms enjoy the least complexity, they can be used when traffic load is not too high.
TABLE IV: Throughput Comparison

| Algorithm | JADE | DASE | DATE |
|-----------|------|------|------|
| Maximum throughput degradation against Type 0 | 0.9% | 2.8% | 2.5% |
| Maximum throughput gain over LEAP | 9.9% | 8.9% | 11.0% |

Fig. 2: Packet throughput of the proposed three scheduling algorithms with type-1 FDRA, as well as LEAP in [3] and an RBG-level type-0 FDRA.

V. CONCLUSION

We have proposed three practical multi-UE scheduling algorithms with type-1 FDRA, i.e., JADE, DASE, and DATE, and compared their performance with each other and with an optimal non-contiguous RBG-level FDRA method and a typical contiguous FDRA algorithm LEAP in the industry. Numerical results demonstrate that JADE can achieve near-optimal performance and outperform LEAP in terms of QoS requirements while having low computational complexity. Additionally, DASE and DATE perform similarly to JADE for small to moderate numbers of UEs, but with substantially lower computational complexity, thus can be adopted in practice under low traffic load conditions. The proposed algorithms are applicable to both downlink and uplink. This work can be extended by considering coarser frequency granularities in contiguous FDRA to mitigate RIV overhead [14].

REFERENCES

[1] 3GPP TS 38.214, V16.2.0, “NR; physical layer procedures for data,” Jun. 2020.
[2] 3GPP TS 38.331, V16.1.0, “NR; radio resource control protocol specification,” Jul. 2020.
[3] I. C. Wong, O. F. Oteri, and J. W. McCoy, “Resource allocation in multi data stream communication link,” U.S. Patent 7 911 934, Mar. 2011.
[4] M. Zhang and Y. Zhu, “An enhanced greedy resource allocation algorithm for localized sc-fdma systems,” IEEE Communications Letters, vol. 17, no. 7, pp. 1479–1482, 2013.
[5] S. Ciccolo and V. Tralli, “Adaptive resource allocation with proportional rate constraints for uplink sc-fdma systems,” IEEE Communications Letters, vol. 18, no. 8, pp. 1419–1422, 2014.
[6] T. Maciel et al., “Method and apparatus for resource allocation to maximize the total data rate in sc-fdma uplink,” U.S. Patent 9 961 682, May 2018.
[7] O. Arafat and M. A. Gregory, “A weight based channel scheduling for LTE-a femtocells,” in 2019 29th International Telecommunication Networks and Applications Conference (ITNAC), 2019, pp. 1–5.
[8] E. E. Tsiropoulou, A. Kapoukakis and S. Papavassiliou, “Uplink resource allocation in sc-fdma wireless networks: A survey and taxonomy,” Computer Networks, vol. 96, pp. 1–28, 2016.
[9] D. Tse, “Forward link multiuser diversity through proportional fair scheduling,” presentation at Bell Labs, Aug. 1999.
[10] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, “Providing quality of service over a shared wireless link,” IEEE Communications Magazine, vol. 39, no. 2, pp. 150–154, 2001.
[11] S. Sun, S. Moon, and J. Fwu, “Practical link adaptation algorithm with power density offsets for 5g uplink channels,” IEEE Wireless Communications Letters, vol. 9, no. 6, pp. 851–855, 2020.
[12] M. Bennis, M. Debbarh, and H. V. Poor, “Ultra-reliable and low-latency wireless communication: Tail, risk, and scale,” Proceedings of the IEEE, vol. 106, no. 10, pp. 1834–1853, 2018.
[13] 3GPP TR 38.824, V16.0.0, “Study on physical layer enhancements for nr ultra-reliable and low latency case,” Mar. 2019.
[14] K. Kittichokechai, M. Andersson, R. Baldemair, and Y. Blankenship, “Improved frequency domain resource allocation indication,” in 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), 2020, pp. 1–5.