Resource Optimization with 5G Configured Grant Scheduling for Real-Time Applications

Yungang Pan, Rouhollah Mahfouzi, Soheil Samii, Petru Eles and Zebo Peng
Embedded Systems Lab, Linköping University, Sweden
{yungang.pan, rouhollah.mahfouzi, soheil.samii, petru.eles, zebo.peng}@liu.se

***Abstract*—5G is expected to support ultra-reliable low latency communication to enable real-time applications such as industrial automation and control. 5G configured grant (CG) scheduling features a pre-allocated periodicity-based scheduling approach which reduces control signaling time and guarantees service quality. Although this enables 5G to support hard real-time periodic traffics, efficiently synthesizing the schedule and achieving high resource efficiency while serving multiple traffics, is still an open problem. To address this problem, we first formulate it using satisfiability modulo theories (SMT) so that an SMT-solver can be used to generate optimal solutions. For enhancing scalability, two efficient heuristic approaches are proposed. The experiments demonstrate the effectiveness and scalability of the proposed technique.**

***Index Terms*—5G, deterministic periodic traffic, configured grant scheduling, resource optimization**

I. Introduction

5G has been designed to support a wide range of applications including ultra-reliable low latency communication, which is one of the new and vital features of 5G in supporting real-time applications and services required by, e.g., industrial automation and manufacturing [1]. In many industrial applications, deterministic periodic traffic is one of the most common traffic types [2], which would generate data packets periodically with extremely strict timing constraints.

To facilitate the scheduling of deterministic periodic traffic flows (TFs), 5G introduces the CG scheduling scheme to assign dedicated resources for data transmissions based on TF requirements [3]. Compared to the conventional "dynamic scheduling" scheme, tremendous control overhead (a large number of scheduling requests/scheduling grant messages) can be avoided. A detailed illustration of the CG scheduling scheme depicting the uplink transmission between the user and the base station (BS) is given in Fig. 1. There are two "configurations," each indicating a specific data transmission scheme: the allocated time-frequency resources (gray rectangles in the figure) and the period between two adjacent data transmissions (period 1 and period 2 in the figure). Specifically, configuration 1 is valid for the first three data transmissions, in which the period equals the periodicity of the traffic. After that, configuration 2 is activated for the next three data transmissions with a new time-frequency resource allocation scheme and a shorter period (i.e., period 2) compared to configuration 1. *Generally, more configurations bring a higher level of scheduling flexibility for data transmissions, while at the same time, the associated control overhead transmitting configurations may decrease the wireless resource efficiency.* Therefore, the scheduling of the CG scheduling scheme is not trivial, especially with multiple TFs. An illustrative example demonstrating the above trade-off by three candidate schedules is shown in Fig. 2. From (a) to (b), a much more appropriate time-frequency resource allocation pattern contributes to a higher resource efficiency (i.e., less frequency resource usage). From (b) to (c), one more configuration brings higher scheduling flexibility, thus improving resource efficiency while one more control message is needed.

∗This research has been supported by the Swedish national strategic research program ELLIIT.

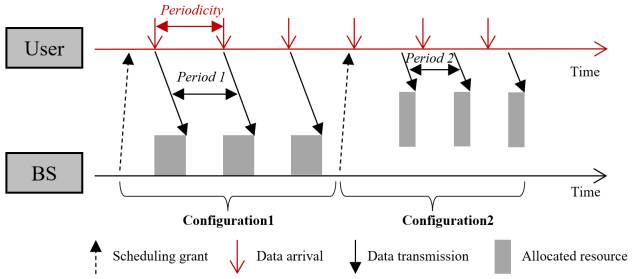


Fig. 1: Illustration of the configured grant scheduling scheme.

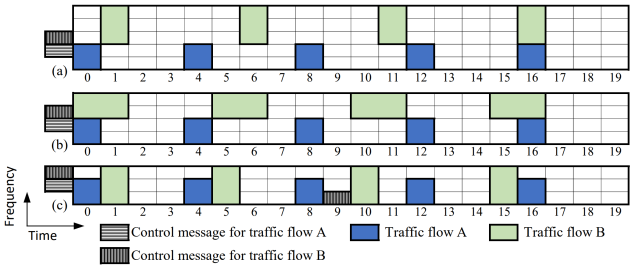


Fig. 2: Illustration of the trade-off between scheduling flexibility and control overhead for two periodic TFs.

There are few closely related studies in this regard [4]–[7]. The impact of explicitly involving additional control messages on resource efficiency has not been investigated to the authors' knowledge. Hence, how the scheduling of data transmission and the involved control overhead impact the overall system performance has been studied in this work. The core problem is: *how to solve the complex multi-traffic resource optimization problem with CG scheduling?* The problem formulation, effective heuristic approaches, and experiments are presented briefly as follows.

II. Problem Formulation

The scheduling problem we solve aims at allocating enough resources within the suitable time windows so as to fulfill the TFs' requirements while minimizing the total number of used frequency resource units called resource blocks (RBs).

978-3-9819263-7-8/DATE23/© 2023 EDAA

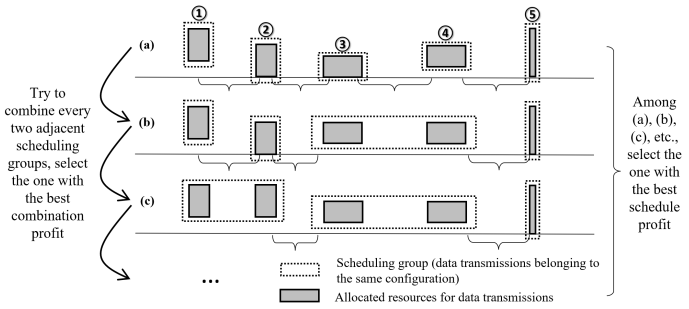


Fig. 3: Illustration of the working process of algorithm CoU.

The inputs include the initial offsets, the transmission periods, the payloads, and the latency requirements of all TFs, while the outputs are the configurations for the transmission of data packets and the corresponding control messages for all TFs. The scheduling window is the same as the hyper period of the transmission periods of all TFs.

III. Solutions

A. SMT-Based Solution

We proposed an SMT-based solution to establish an exact model of the scheduling problem. First, we define the concept of schedule/configurations. Next, we establish all related constraints such as the number of time-frequency resource units required, and the deadlines for all data packets among others. Then, we set up the optimization goal of minimizing the frequency resource usage. Finally, we use the off-the-shelf Z3 solver [8] to solve the SMT model.

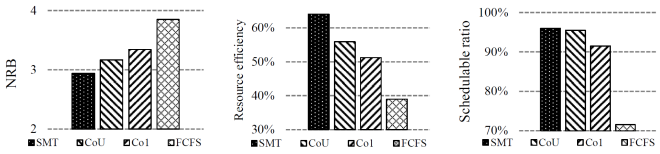
B. Heuristic Approaches

We developed the heuristic approaches Co1 and CoU, since the SMT-based technique suffers from severe scaling problems. The first technique, Co1, follows the basic idea of CG scheduling that all packets are scheduled by one configuration, thus avoiding any control overhead. We establish the Co1 algorithm based on the exhaustive searching scheme, in which four scheduling parameters needed to be figured out: 1) period (time interval between every two adjacent data transmissions); 2) offset (offset for the first data transmission); 3) the number of time slots used for each data transmission; 4) start index and the number of used RBs.

In contrast with Co1, we propose to utilize multiple configurations to enhance scheduling flexibility. Therefore, the involved control messages need to be considered carefully. CoU is such an approach in which an unlimited number of configurations can be used. CoU adopts a step-by-step configuration-reduced approach as shown in Fig. 3. From (a) to (b) then to (c), fewer configurations are needed. Finally, all cases of different numbers of configurations are explored. The intuition of the "combination" is that when we start allocating optimal resources for each data transmission locally, combining two adjacent data transmissions would generally reduce resource efficiency for data transmission yet save resources due to the decrease in the number of control messages.

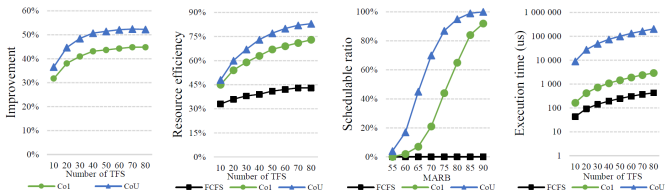
IV. Experiments

Four metrics are used to evaluate the performance of the proposed approaches, including the average number of frequency RBs (NRBs), the resource efficiency, the schedulable ratio,



(a) NRB (b) Resource efficiency (c) Schedulable ratio

Fig. 4: Experimental results of the small scale systems.



(a) NRB improvement compared to FCFS (b) Resource efficiency (c) Schedulable ratio (d) Execution time

Fig. 5: Experimental results of the large scale systems.

and the execution time. A simple heuristic approach "FCFS" following the basic "first come, first served" principle is used as a baseline solution. The maximum allowed number of RBs (MARB) is set for the test of the schedulable ratio. Small scale and large scale experiments with different numbers and characteristics of TFs are conducted. The experimental results are shown in Fig. 4 and Fig. 5. In general, CoU performs close to the optimal SMT-based solution for small scale systems. CoU generally outperforms Co1 and FCFS for large scale systems. Furthermore, we studied the relationship between the algorithm's performance and the characteristics of the test cases including the number of TFs, the latency requirement, the diversity of the transmission period, the payload size, and the control message size. For more details in terms of the implementation and the experiments see https://github.com/dlzrmr99/CGScheduling.

V. Conclusion

This paper addresses the configured grant scheduling problem while considering multiple real-time applications and resource optimization. An SMT-based exact solution and two heuristic solutions are proposed. The experiments demonstrate the performance and superiority of the proposed CoU heuristic approach.

References

[1] "Service requirements for cyber-physical control applications in vertical domains," document 3GPP, Tech. Rep. TS 22.104 V16.5.0, Sep. 2020.

[2] "Study on communication for automation in vertical domains," document 3GPP, Tech. Rep. TR 22.804 V16.3.0, Jul. 2020.

[3] "Nr; nr and ng-ran overall description," document 3GPP, Tech. Rep. TS 38.300 V16.8.0, Dec. 2021.

[4] G. Karadag *et al.*, "Qos-constrained semi-persistent scheduling of machine-type communications in cellular networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 5, pp. 2737–2750, 2019.

[5] J. García-Morales, M. C. Lucas-Estañ, and J. Gozalvez, "Latency-sensitive 5g ran slicing for industry 4.0," *IEEE Access*, vol. 7, pp. 143 139–143 159, 2019.

[6] R. B. Abreu *et al.*, "Scheduling enhancements and performance evaluation of downlink 5g time-sensitive communications," *IEEE Access*, vol. 8, pp. 128 106–128 115, 2020.

[7] N. Jiang, A. Aijaz, and Y. Jin, "Recursive periodicity shifting for semi-persistent scheduling of time-sensitive communication in 5g," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 01–06.

[8] L. d. Moura and N. Bjørner, "Z3: an efficient smt solver," in *International conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2008, pp. 337–340.