Optimal Functional Split Selection and Scheduling Policies in 5G Radio Access Networks

Iordanis Koutsopoulos
Athens University of Economics and Business
Athens, Greece

***Abstract*—We theoretically study the joint functional split selection and scheduling problem that arises in a centralized radio-access-network (C-RAN) architecture with a central location connected to a set of Remote Radio Heads (RRHs) through fronthaul links. A set of LTE frames at the central location need to be transported to RRHs through the fronthaul topology, and one frame is destined towards one Remote Radio Head (RRH). Each frame contains data to be transmitted to users in the corresponding RRH. For each frame, a functional split needs to be selected out of a discrete set of available options, and each option corresponds to a different split of the baseband processing load between the server at the central location and the one residing at the RRH. A functional split also requires that a certain amount of data traverses the fronthaul link to the RRH. Prior to transmission through the fronthaul, frames need to be scheduled at the central location server. The total latency experienced by the frame is the sum of computation, data communication and scheduling delays, and it depends on both the schedule and the functional split selection.**

We seek to characterise the complexity of finding the joint functional split selection and scheduling policy that minimizes the total latency and of the policy that minimizes the maximum latency over all RRHs. The former objective becomes equivalent to a constrained shortest-path problem which is NP-Hard, although the scheduling problem for given functional splits and the functional split selection problem for given scheduling policy are polynomially solvable. The latter objective is also NP-Hard, while the problem of scheduling for given functional splits is optimally solvable through equivalence with single-machine scheduling for maximizing lateness, and the functional split selection problem for fixed schedule is a mixed-integer linear program.

I. Introduction

The centralized radio-access-network (C-RAN) architecture, also referred to as Cloud RAN is expected to leave a firm mark in 5G and beyond 5G radio-access technologies. Fueled by the cheap computational power and advances in cloud computing, C-RAN architectures migrate the computational load of baseband and other physical-layer functionalities from the remote base stations (BSs)–the Remote Radio Heads, RRHs– to a central location with a pool of general-purpose processors which is also called baseband-unit pool (BBU) location. Such functionalities are channel encoding and error correction decoding, modulation and demodulation, resource mapping and demapping, channel estimation and equalization, Fast Fourier transform (FFT) and its inverse, analog-to-digital and digital-to-analog conversion, and antenna radio transmission and reception.

Centralization simpifies the structure of remote BSs, and thus it eases maintenance and reduces deployment costs of dense cellular networks. Further, it facilitates virtualization of computation resources at the BBU location. It also allows the flexible allocation of a pool of radio and computational resources over a large set of cells. Thus, it exploits statistical multiplexing, and it facilitates an array of advanced transmission techniques such as coordinated multi-point (CoMP) transmission and inter-cell interference management which can now be implemented through global optimization over different cells, with the corresponding computations held at the BBU location.

In C-RAN architectures, the transport network from the BBU location to the remote BSs is known as the *fronthaul* network. In traditional RAN architectures, the BBU location did not exist, and all layer functionalities and computations were executed at the remote BSs. Thus, the amount of traffic transported to the BSs and the delay requirements were dictated solely by user requirements in throughput and delay. With C-RAN, the trend moved to the other extreme in which almost all the computation load is carried at the central BBU location, thus leaving the RRH with only time-domain RF processing and A/D functionalities. This approach places significant pressure on the fronthaul to deliver a large amount of traffic comprising baseband samples under very stringent delay constraints.

The recently proposed concept of flexible functional splits seeks to strike a compromise between the two extremes above in the sense that different baseband functions may be allocated at the BBU and the RRH locations. A functional split is an option for partitioning the chain of baseband processing functions into two sub-chains of functions, one of which is executed at the BBU and the other at the RRH.

A certain choice of functional split implies a certain amount of *computational load* for the BBU because of the sub-chain of functions assigned to the BBU, as well as an amount of computational load for the RRH. Different functional splits result in different amounts of *transported traffic* from the BBU to the RRH over the fronthaul topology. Such data is further used by the RRH to execute the rest of the computation chain and transmit to the user. For example, consider the choices

where the modulation function in the baseband processing chain of a user may be performed at the BBU or at the RRH in the downlink. Assume an M-QAM modulation scheme and let L be the number of bits needed per complex baseband sample. Consider a single radio resource block (RRB) in LTE-A containing a user symbols. In a functional split where the modulation is the first function in the chain of computations performed at the RRH, $a \log_2 M$ bits need to be conveyed from the BBU to the RRH so that the RRH continues with the rest of the baseband chain until antenna transmission. On the other hand, in a functional split where the modulation is the last function in the chain of computations performed at the BBU, $2aL$ bits that represent the I/Q complex baseband samples need to be communicated to the RRH.

In C-RANs, a fundamental performance metric in the downlink is latency, measured as the time elapsed from the moment a request for transmission of a frame in the downlink arises at the BBU pool, until data is ready for transmission at the RRH. Of similar meaning is the latency in the uplink; it is the time elapsed between user signal reception at the RRH antenna and user message reception at the medium access layer of the BBU. Latency depends on the selection of the functional split through the delay associated with the amount of computation at the BBU and RRH locations and through the data transport delay at the fronthaul link. Furthermore, different requests for data transmission compete for the finite computational resources at the BBU location, hence appropriate scheduling of computation requests is needed, since a transmission request experiences additional delay due to the service of other requests in the queue prior to that request.

In order to reap the benefits of centralized C-RAN architectures with flexible functional splits, BBU-to-RRH delays of few msecs are needed. In this paper, we consider functional split selection and scheduling as two prime such mechanisms. The space of possible solutions is very large, since typically there exist a few (e.g. 5-6) functional split options and an exponential number of possible frame orderings in the scheduler, hence a judicious optimization of mechanisms that affect delay is need. A set of LTE frames at the BBU location need to be transported to RRHs, one frame destined towards one RRH. Each frame contains data to be transmitted to users in the RRH. For each frame, a functional split needs to be selected out of a discrete set of options. Prior to transmission through the fronthaul link, frames need to be scheduled for processing at the BBU server. The total latency experienced by the frame is the sum of computation, data communication and scheduling delays, and it depends on both the schedule and the functional split selection. The contribution of the work to the literature is as follows.

- We characterise the complexity of finding the joint functional split selection and scheduling policy that minimizes the total latency and that of the policy that minimizes the maximum latency over all RRHs.
- We show that the former objective becomes equivalent to a constrained minimum-cost path problem which is NP-Hard, although the scheduling problem for given functional splits and the functional split selection problem for given scheduling policy are polynomially solvable.
- We show that the latter objective is also NP-Hard, while the partial problem of scheduling for given functional splits is optimally solvable, and the functional split selection problem for fixed schedule is a mixed-integer linear program.

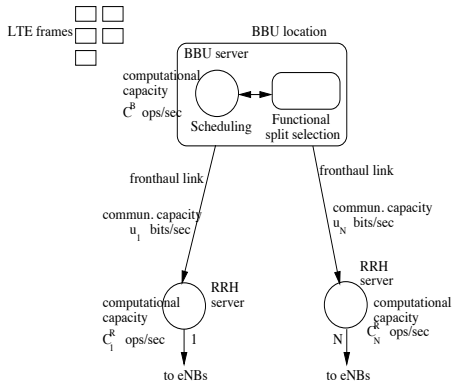

Fig. 1. Depiction of the system topology and basic mechanisms in our model.

To the best of our knowledge the joint consideration of functional splits and BBU server scheduling is novel. The work that is closest to ours is [1]. Compared to this work which adheres to a network calculus approach, we follow an algorithmic approach with objectives that reflect the interplay and sequence of computation and communication in 5G RAN architectures. In addition, we bring forth BBU server scheduling, and we characterize the jointly optimal functional split and scheduling policy. The paper is organized as follows. In section II we show the model and assumptions. In section III we present the formulation and characterize the complexity of the solution for the two objectives above and for the subproblems where we fix the scheduling or the functional split selection policy. Related work is discussed in section IV. We conclude in section V.

II. Model

We consider a cloud-RAN architecture with a central location with a set of cloud (BBU) servers. We assume that BBU servers are colocated, hence they can be considered as a single server with total computational capacity C^B operations per sec.

There also exists a set $\mathcal{R}$ of N RRHs, and each RRH has within its reach a set of users. We consider a scenario with no coordinated multi-point transmission so that a user receives data intended for her only through one RRH. Fix attention to a system snapshot in which a number N_j of LTE frames wait at the BBU location and need to be transported from the BBU location to RRH j, for $j = 1, \ldots, N$. We assume for simplicity that $N_j = 1$ for each j so that we refer to frame j as the frame intended for RRH j. Each RRH j

includes a transmission module and a server of computational capacity C_j^R. We consider the simplest fronthaul topology in which the BBU location is connected to RRH j through a link of communication capacity u_j bits per sec. The topology is depicted in Fig. 1.

An LTE frame j at the BBU location may be abstracted as a job with computational requirements (load) w_j. This load amounts to the total number of operations needed to execute the entire chain of baseband processings steps. The computational requirements of a frame may differ in general, depending on various factors e.g. whether MIMO processing is part of baseband processing.

A discrete set $\mathcal{S} = \{1, 2, \ldots, |S|\}$ of $|\mathcal{S}| = K$ options for functional splits are available. For the LTE frame of each RRH, each functional split specifies a partition of the computation chain into two sub-chains, one of which is executed at the BBU and the other at the RRH. Fig. 2 depicts a chain of computations and a possible split that partitions the chain in two sub-chains. For each LTE frame j a functional split $s \in \mathcal{S}$ corresponds to a pair of values $(w_{j,s}^B, w_{j,s}^R)$ that denote the computational load of RRH frame j that is assigned at the BBU and RRH servers respectively, and such that $w_{j,s}^B + w_{j,s}^R = w_j$. A functional split s corresponds also to a volume of data $r_{j,s}$ to be transported from the BBU to RRH j. Let $s(j) \in \mathcal{S}$ denote the functional split selected for RRH j and let $\mathbf{s} = (s(j) : j \in \mathcal{R})$ the functional split selection policy for all RRHs.

LTE frames wait for transmission at the BBU server. The server chooses a functional split policy $\mathbf{s}$ and a scheduling policy $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$ where $\pi_j \in \{1, \ldots, N\}$ is the order with which the j-th LTE frame is served; e.g. if $\pi_j = 1$ it means that frame j is executed first. The server executes the load $w_{j,s(j)}$ of each frame j according to the order specified in the schedule. We assume that the BBU server devotes its entire computational capacity to the RRH served each time, hence the time needed to serve the load is

$$\tau_{j,s(j)}^B = \frac{w_{j,s(j)}^B}{C^B}\,. \tag{1}$$

After completing the computation for RRH j, the BBU location transmits the $r_{j,s(j)}$ bits over the fronthaul link to RRH j, and this takes time

$$\tau_{j,s(j)}^L = \frac{r_{j,s(j)}}{u_j}\,. \tag{2}$$

When this data arrives at the RRH, the sub-chain of computation allocated to the RRH is executed there in time

$$\tau_{j,s(j)}^R = \frac{w_{j,s(j)}^R}{C_j^R}\,. \tag{3}$$

An RRH frame j experiences a scheduling delay due to the execution of frames prior to j. This delay depends on the scheduling policy and also on the functional splits of frames scheduled before j, and it is equal to

$$T_j(\boldsymbol{\pi}, \mathbf{s}) = \sum_{i:\pi_i < \pi_j} \tau_{i,s(i)}^B\,. \tag{4}$$

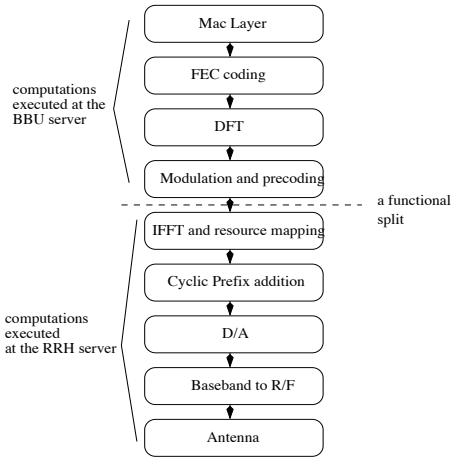

Fig. 2. Typical chain of baseband computations and a possible functional split.

Thus, the total latency experienced by RRH j is

$$D_j(\boldsymbol{\pi}, \mathbf{s}) = T_j(\boldsymbol{\pi}, \mathbf{s}) + \tau_{j,s(j)}^B + \tau_{j,s(j)}^L + \tau_{j,s(j)}^R\,. \tag{5}$$

III. Problem Formulation and Solution

The latency of each RRH j in equation (5) depends on the schedule $\boldsymbol{\pi}$ and the functional split policy $\mathbf{s}$. A given choice of functional splits affects service times $\tau_{j,s(j)}^B$ of each frame j and hence the latencies of all frames scheduled after j.

A first objective, which we refer to as objective I, is to find the scheduling policy $\boldsymbol{\pi}$ and the functional split policy $\mathbf{s}$ so as to minimize total latency. This can be formulated as follows:

$$\min_{\boldsymbol{\pi}, \mathbf{s}} \sum_{j \in \mathcal{R}} D_j(\boldsymbol{\pi}, \mathbf{s})\,, \tag{6}$$

such that $\boldsymbol{\pi} \in \Pi$ and $\mathbf{s} \in \mathcal{S}^N$, where Π is the set of all permutations of $\{1, \ldots, N\}$ and $\mathcal{S}^N$, the set of all functional split policies is the N-times Cartesian product of $\mathcal{S}$.

In order to impose a sense of fairness in treating LTE frames of different RRHs, a second objective (objective II) is to minimize the maximum latency over all RRHs, i.e.

$$\min_{\boldsymbol{\pi}, \mathbf{s}} \max_{j \in \mathcal{R}} D_j(\boldsymbol{\pi}, \mathbf{s})\,, \tag{7}$$

with $\boldsymbol{\pi} \in \Pi$ and $\mathbf{s} \in \mathcal{S}^N$.

A. Minimum-sum-latency problem

First, we consider objective I. For clarity, we first study it for given functional split selection policy and then for given scheduling policy. Finally we solve problem (6).

1) Fixed functional split selection policy: If the functional split selection policy $\mathbf{s}$ is fixed, the scheduling policy that minimizes total latency is the shortest-job-first (SJF) one. That is, RRHs j are scheduled in increasing order of $\tau_{j,s(j)}$.

2) Fixed scheduling policy: If the scheduling policy $\boldsymbol{\pi}$ is fixed, we need to find a functional split allocation policy that minimizes total latency. Let $d_{j,s(j)} = \tau^L_{j,s(j)} + \tau^R_{j,s(j)}$. We use (5) to write the objective (6) as

$$\sum_{j\in\mathcal{R}} D_j(\mathbf{s}) = \sum_{j\in\mathcal{R}} \left(\tau^B_{j,s(j)} + d_{j,s(j)} + \sum_{i:\pi_i<\pi_j} \tau^B_{i,s(i)}\right). \quad (8)$$

By an index exchange argument, we can write

$$\sum_{j\in\mathcal{R}} \sum_{i:\pi_i<\pi_j} \tau^B_{i,s(i)} = \sum_{i\in\mathcal{R}} \sum_{j:\pi_j>\pi_i} \tau^B_{i,s(i)} = \sum_{i\in\mathcal{R}} (N-\pi_i)\tau^B_{i,s(i)}\,. \quad (9)$$

Note that for a scheduled RRH i, $(N-\pi_i)\tau^B_{i,s(i)}$ is the delay incurred by i to RRHs that are scheduled after it in the scheduling policy $\boldsymbol{\pi}$, and therefore

$$\sum_{j\in\mathcal{R}} D_j(\mathbf{s}) = \sum_{j\in\mathcal{R}} \left(d_{j,s(j)} + (N-\pi_j+1)\tau^B_{j,s(j)}\right). \quad (10)$$

Consider a directed acyclic graph (DAG) G with NK nodes, one node (j,s) for RRH frame j in a certain position π_j in the schedule, and each possible functional split $s \in \mathcal{S}$. Then, from each node (j,s) with $\pi_j \leq N-1$, consider outgoing links to nodes (i,s) with $\pi_i = \pi_j + 1$ and $s \in \mathcal{S}$ and for i, i.e. towards each possible functional split of the next RRH scheduled in schedule $\boldsymbol{\pi}$. The weight of this link is defined as

$$\beta_{j,s} = (N-\pi_j+1)\tau^B_{j,s} + d_{j,s}\,. \quad (11)$$

There are also two special nodes f and q, and we connect node f to nodes (j,s) with $\pi_j = 1$ with directed links of weight 0. We also add directed links from node (j,s) with $\pi_j = N$ to node q with links of weight $\tau^B_{j,s} + d_{j,s}$. The graph without link weights is depicted in Fig. 3 for $N=3$ RRHs and $K=2$ possible functional splits.

We observe that a path $\mathcal{P}$ from f to q in graph G corresponds to a selection of functional splits $\mathbf{s} = (s_1,\ldots,s_N)$ for RRHs that are scheduled with policy $\boldsymbol{\pi}$. Further, the sum of costs of the links traversed by a path $\mathcal{P}$ is equal to the total latency $\sum_{j\in\mathcal{R}} D_j(\mathbf{s})$. A path $\mathcal{P}^*$ of minimum cost corresponds to a functional split policy $\mathbf{s}^*$ that minimizes total latency. Therefore the problem of finding the optimal functional split policy is equivalent to finding a min-cost path in a DAG. The min-cost path can emergy through the Dijkstra algorithm in $O(|E| + |V|\log|V|)$ time, where $|V|, |E|$ are the number of nodes and links. In G, there exist $(NK+2)$ nodes and $O(NK^2)$ number of links, thus the algorithm runs in $O(NK^2 + NK\log N)$ time.

3) Joint functional split selection and scheduling policy: In order to study the joint functional split selection and scheduling problem, we define another graph G'. To visualize G', consider nodes placed in N columns, where each column r has NK nodes. Each node is defined through the triad (r,j,s), and in each column r there is one node for each possible RRH j scheduled in position r and each possible functional split s. We add outgoing links from a node (r,j,s) to nodes $(r+1,j',s)$ with $j' \neq j$ and $s \in \mathcal{S}$. That is, outgoing links exist from a node (r,j,s) to all nodes corresponding to the next position, except those nodes that correspond to the same RRH. This is because the same RRH cannot be scheduled more than once. The weight of this link is equal to

$$\beta_{r,j,s} = (N-r+1)\tau^B_{j,s} + d_{j,s}\,. \quad (12)$$

There are also two special nodes f' and q', and we connect node f' to nodes $(1,j,s)$ with directed links of weight 0. We also connect links (N,j,s) to node q' with directed links of weight $\tau^B_{j,s} + d_{j,s}$.

As before, the sum of costs of the links traversed by a path $\mathcal{P}$ is equal to the total latency $\sum_{j\in\mathcal{R}} D_j(\boldsymbol{\pi},\mathbf{s})$. However now, associated with each node (r,j,s) (and all outgoing links from this node), is a "resource" level $\ell_{(r,j,s)} = 1$ that models the number of times RRH j is included in the schedule. A jointly feasible functional split selection and scheduling policy $(\boldsymbol{\pi},\mathbf{s})$ corresponds to a path $\mathcal{P}$ from f' to q' such that for each RRH j, the path traverses a node of the form (r,j,s) exactly once, i.e. no RRH is scheduled more than once. In other words, we add in the min-cost path flow formulation [2, Sec.3.4.1], the constraint $\sum_{(u,v):u=(r,j,f)} g_{u,v}\ell_u = 1$ for each j, where (u,v) are the links, $g_{(u,v)} \in \{1,0\}$ is the flow variable that is 1 if link (u,v) is traversed by the solution path, and $\ell_u = 1$ is the resource level of node u.

A path $\mathcal{P}^*$ from f' to q' of minimum cost corresponds to a joint functional split selection and scheduling policy $(\boldsymbol{\pi}^*,\mathbf{s}^*)$ that minimizes total latency such that each RRH is scheduled exactly once. Therefore the problem is equivalent to finding a *resource-constrained* min-cost path in a graph, which is NP-Hard [3]. There exist several heuristics proposed in the literature for solving the problem, see e.g. [5], [6].

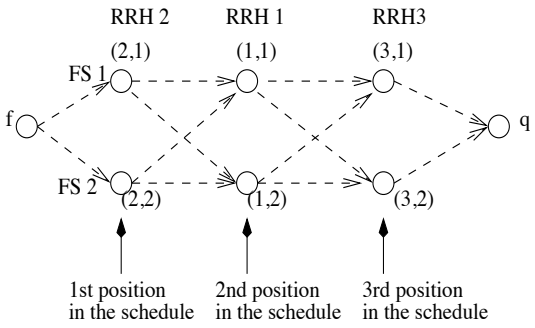

Fig. 3. The graph G for the case of the fixed schedule $\{2,1,3\}$ for $N=3$ RRHs and $K=2$ functional splits. A path from f to q in G corresponds to a functional split selection policy $\mathbf{s}$ while the path cost is equal to the total latency.

B. Min-max latency problem

Objective II aims to balance RRH latencies as much as possible. We can write (7 as

$$\min_{\boldsymbol{\pi},\mathbf{s}} \max_{j\in\mathcal{R}} D_j(\boldsymbol{\pi},\mathbf{s}) = \min_{\boldsymbol{\pi},\mathbf{s}} \max_{j\in\mathcal{R}} \left(T_j(\boldsymbol{\pi},\mathbf{s}) + \tilde{\tau}_{j,s(j)}\right), \quad (13)$$

such that $\boldsymbol{\pi} \in \Pi$, $\mathbf{s} \in \mathcal{S}^N$, with $\tilde{\tau}_{j,s(j)} = \tau^B_{j,s(j)} + \tau^L_{j,s(j)} + \tau^R_{j,s(j)}$.

1) Fixed functional split selection policy: Assume that the functional split selection policy $\mathbf{s}$ is fixed, henceforth we drop $\mathbf{s}$ from the notation. The second term in (13) is a constant and depends only on RRH j, while the first term can be seen as the processing time of the RRH frame at the BBU server. Finding the optimal policy for (13) can be seen to be equivalent to that of scheduling N jobs on a single machine so as to minimize lateness [4]. For a set of N jobs, the *lateness* of a job j under a scheduling policy $\boldsymbol{\pi}$ is defined as $L_j(\boldsymbol{\pi}) = T_j(\boldsymbol{\pi}) - \gamma_j$ where $T_j(\boldsymbol{\pi})$ is the actual job completion time after scheduling policy $\boldsymbol{\pi}$, and γ_j is the deadline by which the job needs to be completed.

The problem of minimizing maximum latency, $L_{max}(\boldsymbol{\pi}) = \max_j L_j(\boldsymbol{\pi})$ is solved by Jackson's earliest-due-date (EDD) rule, according to which jobs j are scheduled in increasing order of deadlines d_j. Our problem is equivalent to the problem of minimizing $L_{max}(\boldsymbol{\pi})$, with $d_j = -\tilde{\tau}_{j,s(j)}$. Thus the optimal scheduling policy for our problem is to schedule RRH frames j in *decreasing* order of $\tilde{\tau}_{j,s(j)}$.

2) Fixed scheduling policy: Now assume that the scheduling policy $\boldsymbol{\pi}$ is fixed. Define binary variables $x_{j,s}$ for $j = 1, \ldots, N$ and $s = 1, \ldots, K$, such that $x_{j,s} = 1$ if functional split s is assigned to RRH j, and 0 otherwise. Let $\mathbf{x} = (x_{j,s} : j = 1, \ldots, N; s = 1, \ldots, K)$. Finding the functional split selection policy that solves optimally (13) is formulated as:

$$\min_{\mathbf{x}} \max_{j \in \mathcal{R}} \big(\sum_{s=1}^{K} \tilde{\tau}_{j,s} x_{j,s} + \sum_{i:\pi_i < \pi_j} \sum_{s=1}^{K} \tau^B_{i,s} x_{i,s} \big) \qquad (14)$$

subject to

$$\sum_{s=1}^{K} x_{j,s} = 1, \, \forall \, j = 1, \ldots, N \, , \qquad (15)$$

with $\mathbf{x} \in \{0,1\}^{NK}$. Define variable

$$p = \max_{j \in \mathcal{R}} \big(\sum_{s=1}^{K} \tilde{\tau}_{j,s} x_{j,s} + \sum_{i:\pi_i < \pi_j} \sum_{s=1}^{K} \tau^B_{i,s} x_{i,s} \big) \, . \qquad (16)$$

Then the problem becomes,

$$\min_{p, \mathbf{x}} p \qquad (17)$$

subject to:

$$\sum_{s=1}^{K} \tilde{\tau}_{j,s} x_{j,s} + \sum_{i:\pi_i < \pi_j} \sum_{s=1}^{K} \tau^B_{i,s} x_{i,s} \leq p \, , \forall \, j \, , \qquad (18)$$

and such that $\sum_{s=1}^{K} x_{j,s} = 1$ for $j = 1, \ldots, N$, and $\mathbf{x} \in \{0,1\}^{NK}$, and real $p \in \mathcal{R}^+$. This is a mixed-integer-linear program (MILP) and can be solved with relaxation methods to a linear program, Lagrangian duality [2, Ch.10] or with other numerical methods.

3) Joint functional split selection and scheduling policy: We now turn to the joint functional split selection and scheduling policy that solves (13). Consider a simpler hypothetical instance of the problem, call it II', where $\tilde{\tau}_{j,s(j)} = \tau$ for all j. That is, the functional split selection does not affect the second term in the minmax objective (13).

The problem becomes equivalent to a joint job scheduling and processing time selection (out of a discrete set of processing time options) in a single machine so as to minimize the maximum lateness of jobs that have a common completion deadline τ. In [23, Th. 2], the authors refer to this problem as P4', and they show that an instance of a Knapsack problem can be reduced to P4', hence P4' is NP-Hard. So is the case with instance II' and therefore with objective II. A heuristic algorithm that iterates between solving the two sub-problems above could be used to derive a solution to the joint problem.

IV. Related work

In C-RAN architectures, the issue of data transport between the BBU and RRH arises. A first class of data transport protocols currently under consideration are the Common Public Radio Interface (CPRI) [7], [8] and the Open Base Station Architecture Initiative (OBSAI)[9] ones. CPRI transports Constant-Bit-Rate (CBR) raw I/Q sample data over a dedicated channel, while OBSAI uses a packet-based mode to again transmit I/Q samples between the BBU and the RRH, and both use the Radio over Fiber (RoF) solution. This creates tremendous needs for fronthaul bandwidth. The trend to rethink fronthaul transport led to the creation of the IEEE working group on Next Generation Fronthaul Interface (NGFI). NGFI advocates the use of Ethernet for fronthaul data transport as an off-the-shelf alternative that will aid in reaping the benefits of statistical multiplexing and packet routing [10]. Some attempts to build prototypes based on NGFI are recently reported, such as [11]. A state-of-the-art overview of architectures and issues in C-RAN architectures is presented in [12].

A survey of different choices of functional splits and the corresponding requirements for the fronthaul is provided in [13] and [14]. The latter work also includes a discussion on fronthaul technologies and converged fronthaul / backhaul approaches. Several works study the performance of different functional splits in terms of alleviating high data rates between the BBU and RRH locations, such as [15]. The work [1] derives a multi-objective optimization problem for joint fronthaul/backhaul optimization through a network calculus approach. The work in [16] casts the problem of baseband function splitting and placement as a graph clustering problem. The chain of baseband functions is modeled as a directed acyclic graph where each graph node represents a baseband function, while a link between two nodes denotes successive functions. Node weights model the computational costs of carrying out the specific function, and link weights capture the amount of communication traffic that needs to be transported from one function to the next one.

In [17], the authors study the impact of functional split and traffic packetization on fronthaul performance in terms of delay in an NGFI transport scenario. The paper includes numerical studies on the scenario of RRH multiplexing with a view towards choosing the packetization mode and functional split so that a large number of RRHs are supported subject to deadline constraints imposed by the retrasmission protocol.

Another thread of works studies resource management problems in C-RAN architectures. The seminal work [18] studies the problem of optimal partitioning of the set of BSs into subsets and of scheduling computational loads of BSs. Optimality is sought in the sense of maximizing the number of scheduleable BSs subject to deadline constraints for accomplishing the tasks. The work in [19] addresses fronthaul resource allocation issues pertaining to transmission coordination to a user, whereby multiple RRHs jointly transmit to a user in a coordinated fashion. The work [20] studies the problem of virtualized BBU placement with the aim to minimize fronthaul energy consumption. In [21], BBU placement is studied jointly with routing between the BBU to the RRH with the goal of minimizing the deployment cost subject to flow conservation equations. The cost amounts to the number of times when functionalities are split between a BBU and an RRH, plus a cost for the fibre connecting BBUs to RRHs.

Finally, there exists a vast body of literature in operations research on scheduling problems on a single machine whose results are useful in order to understand and characterize the solution of resource allocation problems in our setting (see [4] and references therein). The choice among multiple functional splits for the RRH frame resembles the situation of having controllable processing times for the jobs to be scheduled in the machine [22], [23], [24] with a discrete or continuous set of processing times.

V. Conclusion

We studied theoretically the joint problem of scheduling and functional split selection for minimizing the sum of latencies and for minimizing the maximum latency over RRHs. It turns out that both are computationally hard to solve although the sub-problems when one fixes the scheduling or functional split selection policy and solves over the other may be polynomially solvable. We adhered to an offline approach whereby the set of RRHs are available at the BBU location and we characterized the optimal policy. Various model extensions could be studied, e.g one with multiple BBU servers at the BBU location; then RRH frame assignment to a server should also be considered. More composite fronthaul topologies can also be studied. An interesting extension would be to consider the dynamic case in which RRH frame requests for transmission arise dynamically and queues are formed at the BBU server, at the fronthaul link before data transmission, and at the RRH server. In that case new meaningful performance metrics would need to be defined.

References

[1] A. Tzanakaki *et.al.*, "5G Infrastructures Supporting End-User and Operational Services: The 5G-XHaul Architectural Perspective", *Workshop on 5G Architecture (5GArch), in IEEE ICC,* 2016.

[2] L.A. Wolsey, Integer Programming, *Wiley*, 1998.

[3] M. R. Garey and D.S. Johnson. Computers and Intractability: A guide to the theory of NP-Completeness, *Freeman,* 1979.

[4] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnoy Kan and D.B. Shmoys, "Ch 9. Sequencing and scheduling: Algorithms and complexity", *in Handbook of Oper. Res. and Manag. Sci., Elsevier* vol.4, pp.445-522, 1993.

[5] A. Juttner, B. Szviatovszki, I. Mecs, and Z. Rajko, "Lagrange Relaxation Based Method for the QoS Routing Problem", *Proc. IEEE INFOCOM,* 2001.

[6] N. Boland, J. Dethridge and I. Dumitrescu, "Accelerated label setting algorithms for the elementary resource constrained shortest path problem", Operations Research Letters, vol.34, no.1, pp.58-68, 2006.

[7] CPRI, http://www.cpri.info/.

[8] A. de la Oliva, J.A. Hernandez, D. Larrabeite and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN based LTE scenarios", *IEEE Comm. Mag.* vol.54, no.2, pp.152-159, Feb. 2016.

[9] OBSAI, http://www.obsai.com/.

[10] China Mobile Research, Alcatel-Lucent, Nokia Networks, ZTE Corp., Broadcom Corp., Intel China, "White paper of Next Generation Fronthaul Interface", v1.0, June 2015.

[11] R. Knopp, N. Nikaein, C. Bonnet, F. Kaltenberger, A. Ksentini, R. Gupta, "Prototyping of Next Generation Fronthaul Interfaces (NGFI) using OpenAirInterface", White Paper, EURECOM.

[12] O. Simeone, A. Maeder, M. Peng, O. Sahin and W. Yu, "Cloud Radio Access Network: Virtualizing Wireless Access for Dense Heterogeneous Systems", *J. Commun. Networks,* vol. 18, no. 2, pp.135-149, April 2016.

[13] A. Maeder, M. Lalam, A. De Domenico, E. Pateromichelakis, D. Wuebben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a Flexible Functional Split for Cloud-RAN Networks", *in Proc. EuCNC,* 2014.

[14] J. Bartelt, P. Rost, D. Wuebben, J. Lessmann, B. Melis and G. Fettweis, "Fronthaul and Backhaul requiremements of flexibly centralized radio access networks", *IEEE Wireless Comm. Mag.,* vol.22, no.5, pp.105-111, Oct. 2015.

[15] J. Duan, X. Lagrange and F. Guilloud, "Performance Analysis of Several Functional Splits in C-RAN", *in Proc. Veh. Tech. Conf.,* 2016, Spring.

[16] J. Liu, S. Zhou, J. Gong, Z. Niu and S. Xu, "Graph-based Framework for Flexible Baseband Function Splitting and Placement in C-RAN", *in Proc. ICC*, 2015.

[17] C.-Y. Chang, R. Schiavi, N. Nikaein, T. Spyropoulos, and C. Bonnet, "Impact of Packetization and Functional Split on C-RAN Fronthaul Performance", *in Proc. IEEE ICC,* 2016.

[18] S. Bhaumik, S. P. Chandrabose, M. K. Jataprolu, G. Kumar, A. Muralidhar, P. Polakos, V. Srinivasan, and T. Woo, "A Framework for Processing Base Stations in a Data Center", *in Proc. ACM Mobicom,* 2012.

[19] M. Peng, C. Wang, V. Lau and H. V. Poor, "Fronthaul-constrained cloud radio access networks: insights and challenges", *IEEE Wireless Comm. Mag.,* vol.22, no.2, pp.152-160, Apr. 2015.

[20] R. Riggio, D. Harutyunyan, A. Bradai, S. Kuklinski and T. Ahmed, "SWAN: BaseBand Units Placement over Reconfigurable Wireless Fronthauls", *in Proc. IEEE/IFIP Conf. on Network and Serv. Manag.,* 2016.

[21] F. Musumeci, C. Bellanzon, N. Carapellese, M. Tornatore, A. Pattavina, and S. Gosselin, "Optimal BBU Placement for 5G C-RAN Deployment Over WDM Aggregation Networks", *J. of Lightwave Tech.,* vol.34, no.8, pp. 1963-1970, Apr. 2016.

[22] D. Shabtay and G. Steiner, "A survey of scheduling with controllable processing times", *Discrete Appl. Math.,* vol. 155, pp.1643-1666, 2007.

[23] Z.-L. Chen, Q. Lu and G. Tang, "Single machine scheduling with discretely controllable processing times", *Oper. Res. Letters,* vol.21, pp.69-76, 1997.

[24] E. Nowicki and S. Zdrzalka, "A survey of results for sequencing problems with controllable processing times", *Disc. Appl. Math.,* v.26, pp.271-287, 1990.