applied sciences

MDPI

Article

Latency-Aware DU/CU Placement in Convergent Packet-Based 5G Fronthaul Transport Networks

Mirosław Klinkowski

National Institute of Telecommunications, 1 Szachowa Street, 04-894 Warsaw, Poland; M.Klinkowski@il-pib.pl

Received: 29 September 2020; Accepted: 18 October 2020; Published: 22 October 2020

Abstract: The 5th generation mobile networks (5G) based on virtualized and centralized radio access networks will require cost-effective and flexible solutions for satisfying high-throughput and latency requirements. The next generation fronthaul interface (NGFI) architecture is one of the main candidates to achieve it. In the NGFI architecture, baseband processing is split and performed in radio (RU), distributed (DU), and central (CU) units. The mentioned entities are virtualized and performed on general-purpose processors forming a processing pool (PP) facility. Given that the location of PPs may be spread over the network and the PPs have limited capacity, it leads to the optimization problem concerning the placement of DUs and CUs. In the NGFI network scenario, the radio data between the RU, DU, CU, and a data center (DC)—in which the traffic is aggregated—are transmitted in the form of packets over a convergent packet-switched network. Because the packet transmission is nondeterministic, special attention should be put on ensuring the appropriate quality of service (QoS) levels for the latency-sensitive traffic flows. In this paper, we address the latency-aware DU and CU placement (LDCP) problem in NGFI. LDCP concerns the placement of DU/CU entities in PP nodes for a given set of demands assuming the QoS requirements of traffic flows that are related to their latency. To this end, we make use of mixed integer linear programming (MILP) in order to formulate the LDCP optimization problem and to solve it. To assure that the latency requirements are satisfied, we apply a reliable latency model, which is included in the MILP model as a set of constraints. To assess the effectiveness of the MILP method and analyze the network performance, we run a broad set of experiments in different network scenarios.

Keywords: 5G networks; next generation fronthaul interface; centralized radio access network; packet-switched fronthaul network; resource placement; network optimization; MILP modeling

1. Introduction

The deployment of the 5th generation mobile networks (5G) will lead to a revolutionary transformation of telecommunication networks [1]. By enabling access to new wireless services, including enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable and low-latency communications (URLLC), 5G networks will have a profound impact on different aspects of people's activity [2]. Provisioning of these services in 5G networks will require implementation of centralized and virtualized radio access network architectures [3]. In a centralized radio access network (C-RAN), baseband processing of radio frequency (RF) signals, referred to as a baseband unit (BBU), is separated from the base station/antenna site and moved to a central location, at which the BBUs from different sites are supported. Concurrently, in a virtualized radio access network (vRANs), the BBU processing functions are virtualized and performed on general-purpose processors in a cloud environment.

5G C-RANs will require cost-effective solutions in order to assure connectivity between a large number of antenna sites and centralized BBUs in the so-called fronthaul network [4]. Since current fronthaul technologies are not scalable and flexible in a sufficient way to meet the requirements of

5G services, the research community and industry have worked on developing suitable fronthaul solutions for 5G [5,6]. Among them, the IEEE P1914.1 standard for packet-based fronthaul transport networks [7], which defines the next generation fronthaul interface (NGFI) architecture, is a very most promising one. IEEE has also proposed the IEEE 802.1CM standard specifying time-sensitive networking (TSN) for packet-switched fronthaul networks [8]. In the IEEE 802.1CM standard, Ethernet technology supported by deterministic TSN features is considered for fronthaul networks. In this paper, we focus on the network scenario implementing both mentioned IEEE standards, which we refer to as the NGFI network (see Figure 1).

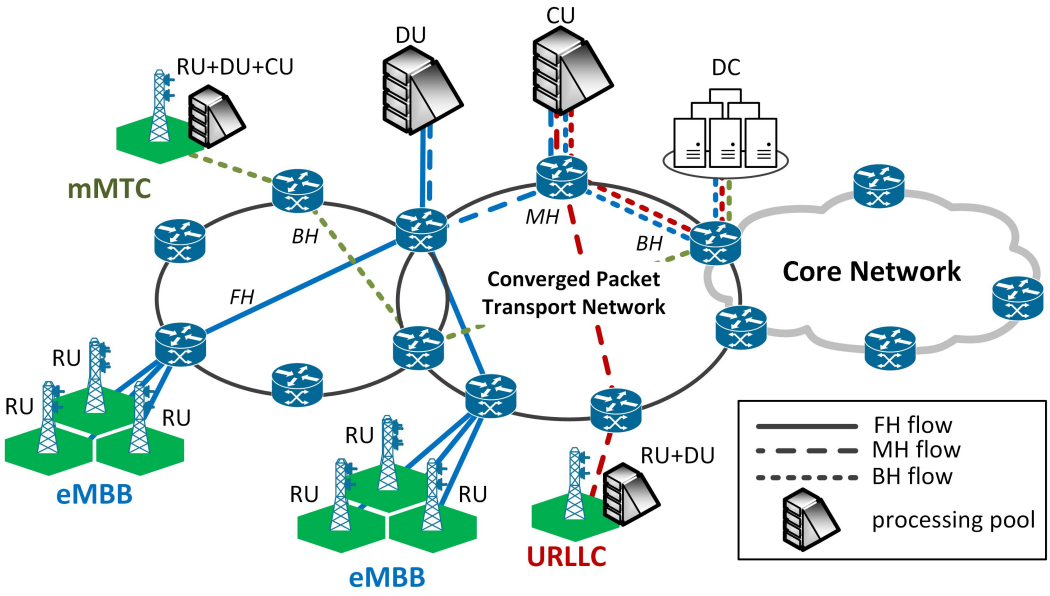

Figure 1. 5G network implementing packet-switched next generation fronthaul interface (NGFI) network architecture.

In NGFI, the radio frequency processing functions are split and performed in radio (RU), distributed (DU), and central (CU) units. In this architecture, the RUs realize low-level physical functions at the antenna sites, while DUs and CUs perform the BBU functions. The DU and CU entities are virtualized, which allows to perform RF processing on general-purpose processors in a processing pool facility [9]. The radio frequency functions that are time-critical are performed at the DU that is located at a processing pool (PP) node in proximity of the RU. This allows to reduce bandwidth requirements of the traffic carried between the DU and CU, whenever the CU is placed in a different PP node. The radio frequency processing is completed in the CU, from which the resulting IP traffic is sent towards the core network for further 5G core and content processing in a data center. The specification and requirements of the virtualized and centralized (RU-DU-CU) radio access network architecture are currently under development within the open ran (O-RAN) initiative founded by major network operators cooperating in the O-RAN Alliance [10].

In the NGFI architecture, three distinct sections can be distinguished, namely fronthaul (FH), midhaul (MH), and backhaul (BH). Fronthaul is the part of the network between the RU and DU, midhaul spans from DU to CU, and backhaul is between the CU and a data center (DC). In the NGFI scenario considered, the data between the RUs, DUs, CUs, and data centers are encapsulated and transmitted as Ethernet frames (packets)—in this paper, terms packet and frame are used interchangeably. The NGFI network allows for convergent transport of different types of flows, including FH, MH, and BH flows, over a shared network. The frames belonging to particular data flows are switched using Ethernet switches (bridges) and routed over a common packet-switched

transport network towards destination nodes. The switches in the network may be connected using both optical and radio links.

A challenging issue in packet-based NGFI networks is quality of service (QoS) provisioning for the traffic flows that have strict latency requirements. These requirements are related to high requirements of 5G services and stringent latency constraints in baseband processing (DU/CU) in the radio access network. Buffering of packets at switch output ports makes the latencies unpredictable and complicates the provisioning of QoS in convergent packet-switched networks. Therefore, latency-aware placement of computational resources for the purpose of radio processing in selected PPs (namely, the placement of DU/CU entities), and proper handing of resulting traffic flows, which may have diverse latency and bandwidth requirements, is in general a difficult optimization task. We call this resource allocation problem a latency-aware DU and CU placement (LDCP) problem. In order to solve LDCP, dedicated optimization methods are required. These methods should be supported by reliable models estimating flows latencies in the packet-switched network to ensure that flows latencies do not exceed allowable limits.

In this paper, we formulate and analyze the LDCP problem in the NGFI network by means of the mixed-integer linear programming (MILP) approach. To assure that the latencies of flows are within allowable limits, we introduce into the MILP formulation of LDCP a set of constraints that represent the worst-case estimates of flows latencies. To analyze the NGFI network performance in different network scenarios, we solve the MILP model using a general purpose mixed-integer programming solver (CPLEX) [11].

The related works and our main contributions are discussed below.

1.1. Related Works

Centralized and virtualized radio access networks have brought some new problems concerning optimized allocation of radio processing and transmission resources. Most optimization studies have focused on the placement of baseband processing units in the conventional centralized radio access network scenario, in which the entire BBU processing is performed at a central site. Among others, MILP formulations for the BBUs placement problem in the C-RAN, in which dedicated point-to-point (P2P) connections are established over an optical network, have been proposed in [12–14]. The authors of [15] extended the analysis to the network resiliency and energy efficiency context. A heuristic was proposed in [16] for solving the BBU placement problem in survivable C-RANs established over optical networks. A graph-based method as well as a genetic algorithm were proposed for the functional split selection and BBU placement problem [17]. The authors of [18] focused on joint functional split selection and data scheduling in BBUs in a fronthaul network with P2P connections and imposed latency constraints. The optimization problem was modeled as an MILP problem, in which optimization goal was to minimize latencies in the network. A literature survey related to resource allocation problems in C-RANs was presented in [19].

There are not many works in the literature concerning optimization of both packet-based fronthaul networks and the C-RAN architectures with distributed DU and CU processing. MILP modeling was applied in [20,21] for functional split selection in the C-RANs in which P2P links are used in the fronthaul and midhaul sections of the network. These works focused on minimization of the power consumption and bandwidth usage in midhaul. The authors of [22] formulated an MILP optimization problem concerning joint placement of DU and CU entities in the C-RANs connected using optical networks. The MILP model was used to analyze the advantages of distributed RU-DU-CU processing in comparison to the C-RAN architecture. In [23], a C-RAN scenario with flexible splitting of radio processing functions into a number of entities placed at different sites of the optical transport network was studied. To formulate the radio processing function placement problem, a MILP approach was used. The authors of [9] addressed the functional split selection problem jointly with the routing problem in a packet-based (RU-CU) C-RAN network. To solve the problem, two heuristic algorithms were developed. The latencies model applied in [9] does not account for packet buffering. In [24,25],

heuristic algorithms were proposed for the problem of routing with latency and flow scheduling constraints in a packet-switched C-RAN network. Recently, the problem of flow allocation with latency constraints in the NGFI network has been addressed in [26,27]. To model the problem, the MILP approach was used, and to solve it an efficient meta-heuristic algorithm was developed.

1.2. Contributions

In this paper, we formulate and study the LDCP problem, which consists in selecting a subset of PP nodes in which the DU and CU radio processing entities have to be placed, assuming QoS (latency) guarantees for FH, MH, and BH flows carried over a common packet-switched transport network. According to our best knowledge, this problem has not been studied in the literature so far. In the majority of prior works, the radio processing functions placed at different network sites were connected using dedicated optical links [20–23]. For this reason, these works did not account for latency guarantees for traffic flows in a packet-switched network. The authors of [9] studied a packet-switched network with FH and BH flows; however, queuing latencies were not taken into account. In [24,25], dynamic latencies were included into analysis; however, only one type of traffic flow, namely MH flow, was assumed. Eventually, in [26,27], we formulated the flow allocation problem that concerned the assignment of DU nodes to a set of RUs and routing of FH and MH flows in the NGFI network under latency constraints. Still the problem addressed in [26,27] was simplified since BH flows were not included into analysis, and the capacities of processing nodes were assumed to be unbounded.

Given the above, our main contribution concerns modeling and solving the LDCP problem in the NGFI network in which latency-sensitive fronthaul, midhaul, and backhaul flows are jointly carried in a convergent packet-based network. Our particular contributions are as follows:

1. development of an MILP optimization model for latency-aware DU/CU placement;
2. in the MILP model, consideration of three different traffic flows (FH, MH, BH) realized jointly in the NGFI network;
3. in the MILP model, consideration of limited PP processing capacities in the NGFI network;
4. reporting and discussion of results of numerical experiments assessing performance of the MILP optimization model proposed and evaluating NGFI network performance in different scenarios.

We would like to stress that the LDCP-MILP model proposed can be applied practically for solving an essential optimization problem in the NGFI network. This problem concerns the allocation of processing resources for realization of radio processing function in a virtualized and centralized radio access network, which is connected using a packet-switched network. In particular, the LDCP optimization problem appears when planning the placement of distributed and centralized units in the NGFI network. Note that the NGFI network is one of the most promising C-RAN solutions, and it is very likely to be deployed in 5G communication networks. In this work, we show the results of such optimization assuming different network topologies and configurations.

The remainder of the article is structured as follows. In Section 2, the network scenario, traffic model, and latency model considered in this work are presented. In Section 3, the LDCP problem is described. Moreover, an MILP formulation of the problem is proposed. In Section 4, numerical experiments are performed. Finally, we present concluding remarks in Section 5.

2. Network Model

In this paper, we study the 5G radio access network that implements the NGFI architecture, which was defined in [7]. The connectivity in the NGFI network is achieved using a fronthaul network consisting of Ethernet switches [8]. In the following, we discuss in details the assumptions concerning the network, traffic model, and latency model considered in this paper.

2.1. NGFI Network

We study the NGFI network which operates with both double-split and a single-split deployment scenarios defined in [7]. In double-split, the baseband functions are split and realized in a distributed way in RU, DU, CU entities, whereas in single-split, DU is co-located with either RU or CU. In the single-split scenario considered in this paper, the DU entity is co-located with CU. The RUs are located close to the antenna site, and the CU and DU are placed at PP nodes, which are spread over different sites of the network.

As defined by the 3GPP organization [28], several options have been distinguished for performing the split of baseband processing functions. According to the indications presented in [7,29], in this work we assume that the functional split between RU and DU implements Option 7.2, and the function split between DU and CU applies Option 2.

We assume that subsets of RUs may be clustered to enable joint processing for the purpose of multi-cell coordination [30]. Accordingly, the DUs associated with the RUs belonging to a cluster must be placed in the same PP node to enable joint processing.

2.2. Traffic Flows

The network supports three different types of flows, namely fronthaul, midhaul, and backhaul flows. The FH flow corresponds to the radio data transmitted between a RU and a DU, the MH flow carries the radio data between a DU and a CU, and the BH flow is the flow of traffic between a CU and the DC. In this work, we assume that one DC supports all CUs. The flows are realized in two directions, namely in an uplink direction (RU→DU→CU→DC) and in a downlink direction (DC→CU→DU→RU). We consider that traffic flows have diversified latency requirements. In particular, the one-way latency limits are 100 µs, 1 ms, and 2 ms, respectively, for FH, MH, and BH flows. Note that depending on the particular network and service scenario, other latency limits may be applied.

2.3. Packet Transport Network

The transport of data flows between RUs, DUs, and CUs, as well as between CUs and the DC, is achieved by means of a packet-switched network. The packet transport network implements the TSN features defined in [8] with the aim to support the transport of latency-sensitive data. Three classes of traffic of different priorities, namely high priority (HP), medium priority (MP), and low priority (LP), are supported in the network. Fronthaul flows need the lowest latencies and they are served as the HP class. The MP class is assigned to midhaul flaws, which may tolerate higher latencies. Eventually, the backhaul traffic is served with the lowest priority.

Each class of traffic has a dedicated queue at the switch output port. The selection of packets for transmission is performed based on the priority levels of packets, in accordance to the strict priority algorithm defined in [31]. In particular, a packet from a non-empty queue of the highest priority is selected first. For the queued up packets of flows of same priority, the selection may be arbitrary. Moreover, preemption of frames is not allowed in the switches [8]. Therefore, the transmission of a lower-priority packet must be completed before the transmission of a higher-priority packet is allowed.

2.4. Traffic Model

We assume the traffic model that we used in [27], which was developed based on [32]. In this traffic model, the data which are carried by traffic flows over the packet transport network have a constant bit-rate. The data are sent by RUs, DUs, and CUs, periodically, as bursts of Ethernet frames. Each remote unit periodically generates the bursts containing radio data and destined to its DU. After processing in DU, the data are again periodically sent in the form of a burst of Ethernet frames to the central unit, in which the radio processing is completed. Finally, CU sends the IP traffic encapsulated into Ethernet frames towards the DC for further 5G core and content processing.

The bursts are not divided in the network into individual frames, but are switched as entire. The frames have the payload of a fixed size equal to 1500 bytes [8]. Additionally, each frame has 42 bytes of overhead [8,32].

The bit-rates of FH, MH, and BH flows have been estimated according to the model provided in [33]. For evaluation purposes, a radio system consisting of four antennas with MIMO and 100 MHz channels was considered. As discussed in Section 2, we assumed functional split Options 7.2 and 2, respectively, in fronthaul and midhaul. The obtained bit-rates of flows are shown in Table 1. Additionally, in Table 1, we present the size of the burst of Ethernet frames (i.e., number of frames) for particular flows. For more details on the traffic model, refer to [27].

Table 1. Bit-rate and burst size (number of frames) of traffic flows assuming functional split options 7.2 (in fronthaul) and 2 (in midhaul).

Direction	Type of Flow	Flow Bit-Rate (Gbit/s)	Burst Size
Uplink	Fronthaul	9.632	52
	Midhaul	1.111	6
	Backhaul	1.111	6
Downlink	Fronthaul	11.113	60
	Midhaul	1.111	6
	Backhaul	1.111	6

2.5. Latencies Modeling

For modeling of flows latencies in the packet-switched network, we applied the latencies model that we presented in [27]. In general, the model accounts for the main sources of latencies in the network [8], which are

- propagation in links,
- storing and forwarding of frames in switches,
- transmission times of bursts of frames, and
- queuing of frames at output ports of switches.

The first three sources of latency are constant (static) and can be estimated easily. In particular, the propagation delay equals the link length divided by the propagation speed (2×10^5 km/s). The delay of a burst transmission in a link is equal to the burst size (see Table 1) multiplied by the transmission time of the frame, which in turn equals the frame size (1542 bytes) divided by the link bit-rate. As in [8], the store-and-forward delay is assumed to be equal to 5 µs.

To model the non-deterministic (dynamic) latency produced by burst queuing in switches, we applied a reliable estimation of latencies. In particular, we estimated the latencies that may occur in the worst possible case. This worst case corresponds to the queuing delays produced by the bursts of frames that belong to other flows than the flow considered, denoted as Y, that might be selected for transmission at the switch output link before the burst of flow Y [8]. To this end, we divided queuing delays into the following two elements:

- delay produced by the flows of either higher or equal priority (t^{HEP}), and
- delay produced by lower priority flows (t^{LP}).

We assume that delay t^{HEP} is produced by the queued-up bursts that belong to all other flows of either equal or higher priority than the priority of flow Y. Concurrently, delay t^{LP} is produced by the largest burst that belongs to a lower-priority flow. We assume that the interfering flows may arrive from different switch input ports, but they go through the same output port as flow Y.

3. LDCP Problem

In this section, we formulate the latency-aware DU/CU placement problem in NGFI networks. In particular, LDCP concerns jointly:

1. placement (in selected PP nodes) of DU and CU entities realizing baseband processing functions for a set of RU nodes, assuming given constraints on
   - maximum processing capacities of the PP nodes,
   - maximum latencies of the fronthaul, midhaul, and backhaul flows realized over the packet transport network between the RUs, the PP nodes selected (for DU and CU processing), and the DC node, and
2. allocation of bandwidth in network links so that to transport FH, MH, and BH flows, assuming given constraints on links capacities.

We illustrate the DU and CU placement in PP nodes and the resulting traffic flows in Figure 2. The network consists of three RUs, three PPs, and one DC. These nodes are linked to five switches in the transport network. The data from RU_1 are carried as fronthaul flow FH_1 through switch v_1 to node PP_1, where a DU entity is located. After DU processing in PP_1, midhaul flow MH_1 goes through switches v_1 and v_5 to node PP_2, where CU processing is performed. After completing the CU processing, the data are carried as backhaul flow BH_1 through switches v_5, v_3, and v_4 to the DC node, where the flow is terminated. RU_2 and RU_3 are grouped into a cluster and, therefore, their DU entities are located in the same PP node, namely in node PP_3. The data from RU_2 and RU_3 are transported in flows FH_2 and FH_3 to PP_3. Flows FH_2 and FH_3 are routed over switches v_2 and v_3. The CU processing for RU_2 and RU_3 is also performed in node PP_3. Therefore, the midhaul flows are not present in the network for these two RUs. After the DU and CU processing in PP_3, flows BH_2 and BH_3 are carried over nodes v_3 and v_4 to the DC.

In the following, we introduce the notation used in problem formulation. Next, we propose an MILP model for the the LDCP optimization problem.

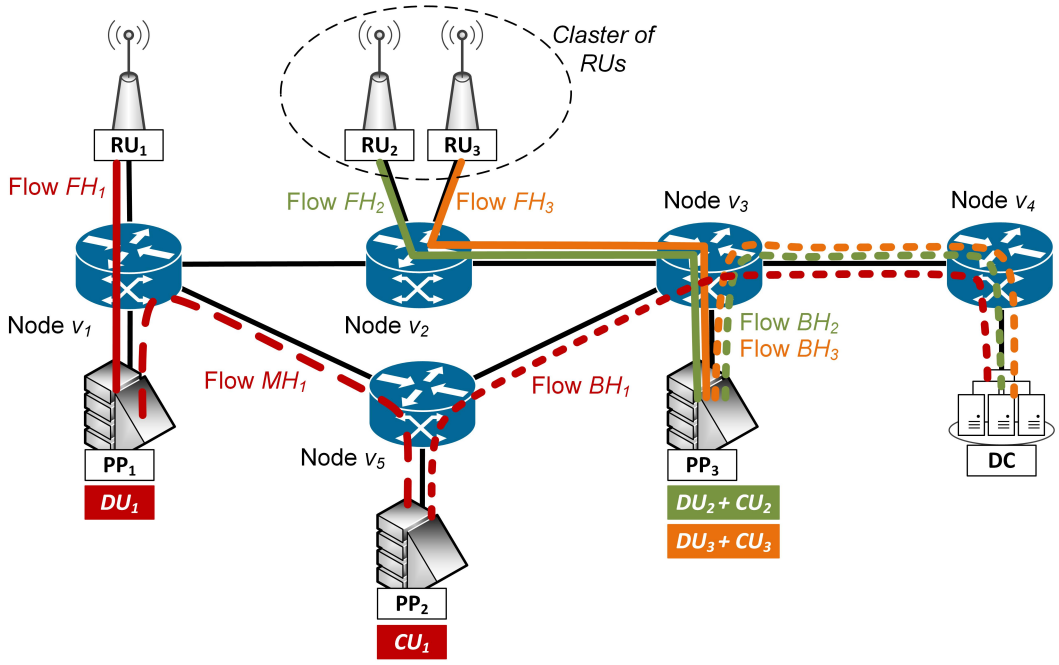

Figure 2. Example of distributed unit (DU) and central unit (CU) placement in processing pool (PP) nodes and resulting traffic flows in the NGFI network.

3.1. Notation

The NGFI network is represented by a directed and connected graph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which $\mathcal{V}$ and $\mathcal{E}$ are the sets of network nodes and links, respectively. Let $\mathcal{V}^R$, $\mathcal{V}^P$, $\mathcal{V}^{DC}$, and $\mathcal{V}^S$ denote the sets of RU, PP, DC, and switching nodes. These sets are disjoint and their sum constitutes the set of all nodes ($\mathcal{V}$). Subgraph $\mathcal{G}^S = (\mathcal{V}^S, \mathcal{E}^S)$ represents the packet transport network, where $\mathcal{E}^S$ is the set of links between the switching nodes ($\mathcal{E}^S \subset \mathcal{E}$). Let $\mathcal{E}^{Sout}$ denote the set of output links of the switches ($\mathcal{E}^{Sout} \subset \mathcal{E}$). The RU, PP, and DC nodes from sets $\mathcal{V}^R$, $\mathcal{V}^P$, and $\mathcal{V}^{DC}$ are connected with some nodes from set $\mathcal{V}^S$ (i.e., with the transport network). Let $K(e)$ denote the capacity of link $e \in \mathcal{E}$ and let $\rho(v)$ be the processing capacity of PP node $v \in \mathcal{V}^P$. Let $L^P(e)$ be the propagation delay in link e. Let $L^{SF}(e)$ be the store-and-forward delay of the switching node that is the origin node of link $e \in \mathcal{E}^{Sout}$. We assume that $L^{SF}(e) = 0$ if link e is not originated in a switch.

The set of clusters is denoted as $\mathcal{C}$. Each cluster $c \in \mathcal{C}$ represents a subset of RUs ($c \subset \mathcal{V}^R$), such that their DUs must be placed and processed together in the same PP node to facilitate multi-cell coordination (see Section 2.1). All RUs belong to some clusters and the clusters are disjoint, i.e., $c_1 \cap c_2 = \emptyset, \forall c_1, c_2 \in \mathcal{C}$.

Let $\mathcal{D}$ be the set of demands. Demand $d \in \mathcal{D}$ is identified with an RU node, and it represents a couple of traffic flows to be realized in the network. The flows are the following:

1. a fronthaul flow—between the RU node and the PP node in which the DU entity is placed;
2. a midhaul flow—between the PP node in which the DU entity is located and a different PP node in which the CU entity is placed. Note that if the DU and CU are located in the same PP node for a given demand, then the MH flow is not present in the network for this demand;
3. a backhaul flow—between the PP node in which the CU entity is located and a DC node.

Let $\mathcal{F}$ denote the set of types of traffic flows, namely $\mathcal{F} = \{\text{FH}, \text{MH}, \text{BH}\}$. Let ρ^D and ρ^C be the processing requirements (loads) of DU and CU entities, respectively.

We assume that for each RU there are two associated demands to be realized in the network, namely an uplink demand and a downlink demand. The uplink demand is realized from RU towards DU, CU, and DC, while the downlink demand has an opposite direction, namely from DC towards CU, DU, and RU. The set of uplink demands is denoted as $\mathcal{D}^U$, and the set of downlink demands is denoted as $\mathcal{D}^D$. These two sets are disjoint, have the same cardinality, and together they form set $\mathcal{D}$, namely, $\mathcal{D} = \mathcal{D}^U \cup \mathcal{D}^D$. The cluster comprising the RU of demand $d \in \mathcal{D}$ is denoted as $C(d)$.

Let $\mathcal{Q}^{HEP}(d, f)$ and $\mathcal{Q}^{LP}(d, f)$ denote the sets of demand-flow pairs $(\bar{d}, \bar{f})$, where $\bar{d} \in \mathcal{D}$ and $\bar{f} \in \mathcal{F}$. Set $\mathcal{Q}^{HEP}(d, f)$ comprises the pairs of either equal or higher priority than flow f of demand d. Set $\mathcal{Q}^{LP}(d, f)$ comprises the pairs of a lower priority than flow f of demand d. For instance, if f is an MH flow, then $\mathcal{Q}^{HEP}(d, f)$ will contain all the FH flows of all demands and the MH flows of all the demands except for d. Concurrently, $\mathcal{Q}^{LP}(d, f)$ will contain all the BH flows of all demands.

We assume that the flows may have different latency and throughput requirements. Let $L^{max}(f)$ denote the maximum one-way latency allowable for flow $f \in \mathcal{F}$. Let $H(d, f)$ denote the bit-rate of flow f of demand d.

Let $\mathcal{V}^{src}(d, f)$ and $\mathcal{V}^{dest}(d, f)$ be the sets of allowable source and destination nodes, respectively, of flow $f \in \mathcal{F}$ of demand $d \in \mathcal{D}$. Assuming that $v^R(d)$ denotes the RU node belonging to demand d, sets $\mathcal{V}^{src}(d, f)$ and $\mathcal{V}^{dest}(d, f)$ are defined as following:

1. if d is an uplink demand, then: $\mathcal{V}^{src}(d, f) = \{v^R(d)\}$ and $\mathcal{V}^{dest}(d, f) = \mathcal{V}^P$ if $f = \{\text{FH}\}$, $\mathcal{V}^{src}(d, f) = \mathcal{V}^P$ and $\mathcal{V}^{dest}(d, f) = \mathcal{V}^P$ if $f = \{\text{MH}\}$, and $\mathcal{V}^{src}(d, f) = \mathcal{V}^P$ and $\mathcal{V}^{dest}(d, f) = \mathcal{V}^{DC}$ if $f = \{\text{BH}\}$;
2. if d is a downlink demand, then: $\mathcal{V}^{src}(d, f) = \mathcal{V}^P$ and $\mathcal{V}^{dest}(d, f) = \{v^R(d)\}$ if $f = \{\text{FH}\}$, $\mathcal{V}^{src}(d, f) = \mathcal{V}^P$ and $\mathcal{V}^{dest}(d, f) = \mathcal{V}^P$ if $f = \{\text{MH}\}$, and $\mathcal{V}^{src}(d, f) = \mathcal{V}^{DC}$ and $\mathcal{V}^{dest}(d, f) = \mathcal{V}^P$ if $f = \{\text{BH}\}$.

For each source–destination pair of nodes in the network, a single path is given. The paths are defined by means of parameter $\alpha(d, f, i, j, e)$, which equals to 1 if flow f of demand d originated in

node $i \in \mathcal{V}^{src}(d,f)$ and terminated in node $j \in \mathcal{V}^{dest}(d,f)$ is routed through link e, and 0 otherwise. Let $L(d,f,e)$ denote the delay produced by transmission of the burst of frames of flow f of demand d at link e.

3.2. Problem Statement

We state the LDCP problem in the following way. Given network topology, traffic demands, routing paths connecting network nodes, capacities of PP nodes and network links, latencies introduced in network elements, and latency limits for flows, we find for all demands a feasible placement of the DU and CU processing entities in the PP nodes under constraints:

1. *Clustering of RUs*: the DUs associated with the RUs that belong to the same cluster are placed in the same PP node;
2. *PP node assignment for DU processing*: for each demand, locate the DU in the PP node that has been assigned to its cluster (i.e., to which its RU belongs to);
3. *PP node selection for CU processing*: a PP node is selected for the CU processing of the demand;
4. *PP node capacity*: the overall DU and CU processing load of all demands processed in each PP node does not exceed the node processing capacity;
5. *Traffic flows*: the traffic flows (FH, MH, and BH) are terminated in the PP nodes in which DU and CU entities are placed; if DU and CU are located in the same PP node, then flow MH is not realized in the network;
6. *Capacity of link*: the overall bit-rates of all flows going through a link must be lower or equal to the link capacity;
7. *Latency of flow*: the latency of a flow cannot be greater than the maximum latency that is allowable for this flow.

The optimization objective considered in this work is to minimize the amount of active PPs and the sum of latencies of all flows in the network. We assume that the former objective is a primary goal and the latter is a secondary goal.

Note that in the problem considered, a single path is provided for each pair of network nodes. Further extensions of the LDCP problem might assume the availability of candidate paths between pairs of network nodes. We will address such a scenario in future work.

3.3. MILP Formulation

As discussed above, the LDCP problem consists in selecting a PP node for DU processing for each cluster, and a PP node for CU processing for each demand. It is allowable to place DU and CU in the same PP node. The latency of flows realized over the network between the PP nodes selected and the source and destination nodes of the demand, using the routing paths given between these nodes, must be kept below the allowable limit. Moreover, the overall processing load in the PP nodes selected and the traffic volume in network links cannot be greater than the available capacity. Hence, binary variable $y_{cv}, c \in \mathcal{C}, v \in \mathcal{V}^D$, indicates whether PP node v is assigned to cluster c for DU processing. There is a pair of binary variables u^D_{dv} and u^C_{dv}, $d \in \mathcal{D}, v \in \mathcal{V}^P$, assigned to each demand, where $u^D_{dv} = 1$ and $u^C_{dv} = 1$ indicate that PP node v realizes DU/CU processing, respectively, for demand d. Besides, binary variable $u^{CD}_{dv}, d \in \mathcal{D}, v \in \mathcal{V}^P$ indicates that both CU and DU of demand d are placed in the same PP node v. Binary variable $y_v, v \in \mathcal{V}^P$, denotes the activation of PP node v. In other words, it is equal to 1 if either DU or CU processing is performed in this node. Binary variable x_{dfij}, where $d \in \mathcal{D}, f \in \mathcal{F}, i \in \mathcal{V}^{src}(d,f), j \in \mathcal{V}^{dest}(d,f)$ indicates that flow f of demand d is realized between nodes i and j. Binary variable x_{dfe}, $d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}$ indicates that flow f of demand d is routed over link e. Binary variable $x_{d\bar{d}f\bar{f}e}$, $d, \bar{d} \in \mathcal{D}$, $f, \bar{f} \in \mathcal{F}$, and $e \in \mathcal{E}$ indicates that flow f of demand d and flow $\bar{f}$ of demand $\bar{d}$ are both carried over link e. Continuous variable $w_{df}, d \in \mathcal{D}, f \in \mathcal{F}$ denotes latency of flow f belonging to demand d. Continuous variables w^{stat}_{dfe} and w^{dyn}_{dfe}, $d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}$ represent, respectively, static and dynamic latency of flow f of demand d in link e. Eventually, continuous

variables w^{HEP}_{dfe} and w^{LP}_{dfe}, $d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}$ denote the latency of flow f of demand d introduced in link e because of either higher-/equal-priority (w^{HEP}_{dfe}) and lower-priority flows (w^{LP}_{dfe}).

The notation used in the MILP model is summarized in Table 2.

Table 2. Notation.

Sets	
$\mathcal{V}$	network nodes
$\mathcal{V}^P$	PP nodes; where $\mathcal{V}^P \subset \mathcal{V}$
$\mathcal{E}$	network links
$\mathcal{E}^{Sout}$	switch output links
$\mathcal{D}$	demands
$\mathcal{D}^U$	uplink demands; $\mathcal{D}^U \subset \mathcal{D}$
$\mathcal{D}^D$	downlink demands; $\mathcal{D}^D \subset \mathcal{D}$
$\mathcal{F}$	types of flows; $\mathcal{F} = \{\text{FH}, \text{MH}, \text{BH}\}$
$\mathcal{Q}^{HEP}(d,f)$	demand-flow pairs of an equal/higher priority than flow f of demand d
$\mathcal{Q}^{LP}(d,f)$	demand-flow pairs of a lower priority than flow f of demand d
$\mathcal{V}^{src}(d,f)$	allowable source nodes of flow f of demand d; $\mathcal{V}^{src}(d,f) \subset \mathcal{V}$
$\mathcal{V}^{dest}(d,f)$	allowable destination nodes of flow f of demand d; $\mathcal{V}^{dest}(d,f) \subset \mathcal{V}$
$\mathcal{C}$	clusters of RUs
Parameters	
$\alpha(d,f,i,j,e)$	$= 1$ if flow f of demand d originated in node i and terminated in node j is routed through link e
$C(d)$	cluster to which the RU of demand d belongs
ρ^D	processing load of a DU
ρ^C	processing load of a CU
$\rho(v)$	processing capacity of PP node $v \in \mathcal{V}^P$
$H(d,f)$	bit-rate of flow f of demand d
$K(e)$	capacity (bit-rate) of link e
$L^P(e)$	propagation delay of link e
$L^{SF}(e)$	store-and-forward delay produced in the origin node (switch) of link e
$L(d,f,e)$	delay produced by transmission of the burst of frames of flow f of demand d at link e
$L^{max}(f)$	maximum one-way latency of flow $f \in \mathcal{F}$
Variables	
x_{dfij}	binary, $x_{dfij} = 1$ if flow f of demand d is realized between nodes i and j
x_{dfe}	binary, $x_{dfe} = 1$ if flow f of demand d is routed over link e
$x_{d\bar{d}f\bar{f}e}$	binary, $x_{d\bar{d}f\bar{f}e} = 1$ if flow f of demand d and flow $\bar{f}$ of demand $\bar{d}$ are routed over link e
u^D_{dv}	binary, $u^D_{dv} = 1$ if DU processing of demand d is performed in PP node v
u^C_{dv}	binary, $u^C_{dv} = 1$ if CU processing of demand d is performed in PP node v
u^{CD}_{dv}	binary, $u^{CD}_{dv} = 1$ if both CU and DU processing of demand d is performed in PP node v
y_{cv}	binary, $y_{cv} = 1$ if cluster c has assigned PP node v for DU processing
y_v	binary, $y_v = 1$ if PP node v is active
w_{df}	continuous, latency of flow f belonging to demand d
w^{stat}_{dfe}	continuous, static latency in link e for flow f belonging to demand d
w^{dyn}_{dfe}	continuous, dynamic latency in link e for flow f belonging to demand d
w^{HEP}_{dfe}	continuous, latency in link e for flow f of demand d caused by higher/equal priority flows
w^{LP}_{dfe}	continuous, latency in link e for flow f of demand d caused by lower priority flows

LDCP-MILP formulation:

$$\text{minimize } z = A \cdot \sum_{v \in \mathcal{V}^P} y_v + \sum_{d \in \mathcal{D}} \sum_{f \in \mathcal{F}} w_{df} \tag{1}$$

where z expresses the number of active PP nodes and total network latency, and A is a weighting coefficient (we assume $A = 10^5$), subject to the constraints:

—RUs clustering—it assures that the DU processing for all RUs belonging to a cluster is performed in the same PP node; in particular, $\forall c \in \mathcal{C}$, the following constraint is imposed:

$$\sum_{v \in \mathcal{V}^P} y_{cv} = 1, \tag{2}$$

—PP node assignment for DU processing—it assures that the DU processing of demands is performed in the PP nodes that have been assigned to the clusters containing the RUs of these demands; in particular, $\forall d \in \mathcal{D}, c = C(d), v \in \mathcal{V}^P$, the following constraint is imposed:

$$u^D_{dv} = y_{cv}, \tag{3}$$

—PP node selection for CU processing—it assures that single PP nodes are assigned for the purpose of CU processing for particular demands; in particular, $\forall d \in \mathcal{D}$, the following constraint is imposed:

$$\sum_{v \in \mathcal{V}^P} u^C_{dv} = 1, \tag{4}$$

—FH and BH flows—it assures that for fronthaul and backhaul flows there is a connection established, respectively, between the RU node and a PP node (in case of FH) and between a PP node and the DC node (in case of BH); in particular, $\forall d \in \mathcal{D}, f \in \{\text{FH}, \text{BH}\}$, we have

$$\sum_{i \in \mathcal{V}^{src}(d,f)} \sum_{j \in \mathcal{V}^{dest}(d,f)} x_{dfij} = 1, \tag{5}$$

—MH flow—assures, for each uplink and downlink demand d, that there is either a MH flow established between a pair of PP nodes (if DU and CU and located in different PP nodes) or such a flow is not realized in the network (if DU and CU are placed in the same PP node); in particular, we have

$$\sum_{j \in \mathcal{V}^{dest}(d,f)} x_{dfvj} + u^{CD}_{dv} = u^D_{dv} \text{ for } d \in \mathcal{D}^U, f = \{\text{MH}\}, v \in \mathcal{V}^{src}(d,f), \tag{6}$$

$$\sum_{j \in \mathcal{V}^{dest}(d,f)} x_{dfvj} + u^{CD}_{dv} = u^C_{dv} \text{ for } d \in \mathcal{D}^D, f = \{\text{MH}\}, v \in \mathcal{V}^{src}(d,f), \tag{7}$$

$$\sum_{i \in \mathcal{V}^{src}(d,f)} x_{dfiv} + u^{CD}_{dv} = u^C_{dv} \text{ for } d \in \mathcal{D}^U, f = \{\text{MH}\}, v \in \mathcal{V}^{dest}(d,f), \tag{8}$$

$$\sum_{i \in \mathcal{V}^{src}(d,f)} x_{dfiv} + u^{CD}_{dv} = u^D_{dv} \text{ for } d \in \mathcal{D}^D, f = \{\text{MH}\}, v \in \mathcal{V}^{dest}(d,f), \tag{9}$$

—Termination of FH and BH flows in the PP nodes selected—it assures that the fronthaul and backhaul flows of demands are terminated in the appropriate PP nodes, namely in which the DU and CU entities of the demands are placed; in particular, the following constraints are imposed:

$$x_{dfiv} = u^D_{dv} \text{ for } d \in \mathcal{D}^U, f = \{\text{FH}\}, i \in \mathcal{V}^{src}(d,f), v \in \mathcal{V}^{dest}(d,f) \tag{10}$$

$$x_{dfvj} = u^C_{dv} \text{ for } d \in \mathcal{D}^U, f = \{\text{BH}\}, v \in \mathcal{V}^{src}(d,f), j \in \mathcal{V}^{dest}(d,f) \tag{11}$$

$$x_{dfvj} = u^D_{dv} \text{ for } d \in \mathcal{D}^D, f = \{\text{FH}\}, v \in \mathcal{V}^{src}(d,f), j \in \mathcal{V}^{dest}(d,f) \tag{12}$$

$$x_{dfiv} = u^C_{dv} \text{ for } d \in \mathcal{D}^D, f = \{\text{BH}\}, i \in \mathcal{V}^{src}(d,f), v \in \mathcal{V}^{dest}(d,f) \tag{13}$$

—Activation of PP nodes—it assures that the PP nodes are active when there are DU/CU entities placed in these nodes; in particular, $\forall v \in \mathcal{V}^P, d \in \mathcal{D}$, the following constraints are imposed:

$$u^D_{dv} \leq y_v, \tag{14}$$

$$u^C_{dv} \leq y_v, \tag{15}$$

—Capacity of PP nodes—it assures that the overall DU and CU processing load in PP nodes does not exceed the capacity of these nodes; in particular, $\forall v \in \mathcal{V}^P$, the following constraint is imposed:

$$\sum_{d \in \mathcal{D}} \left(\rho^D \cdot u^D_{dv} + \rho^C \cdot u^C_{dv} \right) \leq \rho(v), \tag{16}$$

—Capacity of links—it assures that the volume of traffic in network links is not greater than the capacity of links; in particular, $\forall e \in \mathcal{E}$, the following constraint is imposed:

$$\sum_{d \in \mathcal{D}} \sum_{f \in \mathcal{F}} \sum_{i \in \mathcal{V}^{src}(d,f)} \sum_{j \in \mathcal{V}^{dest}(d,f)} H(d,f) \cdot \alpha(d,f,i,j,e) \cdot x_{dfij} \leq K(e), \tag{17}$$

—Utilization of links—it allows to determine whether flows are carried over particular links; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}$, the following constraint is imposed:

$$\sum_{i \in \mathcal{V}^{src}(d,f)} \sum_{j \in \mathcal{V}^{dest}(d,f)} \alpha(d,f,i,j,e) \cdot x_{dfij} = x_{dfe}, \tag{18}$$

—Interfering of flows—it allows to determine whether two different flows use the same switch output link; namely, $x_{d\bar{d}f\bar{f}e}$ is equal to 1 if and only if both flow f of demand d and flow $\bar{f}$ of demand $\bar{d}$ use link e; in particular, $\forall e \in \mathcal{E}^{Sout}$, and $d, \bar{d} \in \mathcal{D}$, $f, \bar{f} \in \mathcal{F}$, except for $d = \bar{d}, f = \bar{f}$, the following constraints are imposed:

$$x_{d\bar{d}f\bar{f}e} \leq x_{dfe}, \tag{19}$$

$$x_{d\bar{d}f\bar{f}e} \leq x_{\bar{d}\bar{f}e}, \tag{20}$$

$$x_{d\bar{d}f\bar{f}e} \geq x_{dfe} + x_{\bar{d}\bar{f}e} - 1, \tag{21}$$

$$x_{d\bar{d}f\bar{f}e} = x_{\bar{d}d\bar{f}fe}, \tag{22}$$

—Dynamic latencies of flows because of the flows of equal/higher priority—it estimates worst-case latencies of flows in the output links of switches caused by either equal- or higher-priority flows; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}^{Sout}$, the following constraint is imposed:

$$\sum_{(\bar{d},\bar{f}) \in \mathcal{Q}^{HEP}(d,f)} x_{d\bar{d}f\bar{f}e} \cdot L(\bar{d},\bar{f},e) = w^{HEP}_{dfe}, \tag{23}$$

—Dynamic latencies of flows because of the flows of lower priority—it estimates worst-case latencies of flows in the output links of switches caused by lower-priority flows; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}, (\bar{d},\bar{f}) \in \mathcal{Q}^{LP}(d,f), e \in \mathcal{E}^{Sout}$, the following constraint is imposed:

$$x_{d\bar{d}f\bar{f}e} \cdot L(\bar{d},\bar{f},e) \leq w^{LP}_{dfe}, \tag{24}$$

—Dynamic latencies of flows—it estimates worst-case latencies of flows produced in the output links of switches; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}^{Sout}$, the following constraint is imposed:

$$w^{HEP}_{dfe} + w^{LP}_{dfe} = w^{dyn}_{dfe}, \tag{25}$$

—Static latencies of flows—it estimates the latencies of flows produced in a network link as the sum of link propagation delay, store-and-forward delay produced in the origin node of the link (if the node is a switch) and burst transmission delay; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}, e \in \mathcal{E}$, the following constraint is imposed:

$$x_{dfe} \cdot \left(L^P(e) + L^{SF}(e) + L(d,f,e) \right) = w^{stat}_{dfe}, \tag{26}$$

—Latencies of flows—it estimates the latencies of flows as the sum of static and dynamic latencies produced in the network links over which the flows are routed; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}$, the following constraint is imposed:

$$\sum_{e \in \mathcal{E}} (w_{dfe}^{stat} + w_{dfe}^{dyn}) = w_{df}, \tag{27}$$

—Maximum latencies of flows—it assures that the latency levels of fronthaul, midhaul, and backhaul flows are within allowable limits; in particular, $\forall d \in \mathcal{D}, f \in \mathcal{F}$, the following constraint is imposed:

$$w_{df} \leq L^{max}(f). \tag{28}$$

The LDCP problem is $\mathcal{NP}$-complete. In particular, it contains constraints (16)–(17) representing the 0-1 knapsack problem, which is $\mathcal{NP}$-complete itself [34]. In Section 4, we investigate the complexity of the LDCP-MILP model using numerical experiments. Afterwards, we use the model to analyze the NGFI network considered.

4. Numerical Results

The LDCP-MILP model is evaluated by means of numerical experiments performed in three networks of different size. The network scenario assumptions discussed in Section 2 are applied. The following topologies of the packet-switched transport network are considered: a 10-node ring network (RING-10), a 16-node double-ring network (DRING-16), and a 20-node mesh network (MESH-20), presented in Figure 3. The topologies have been selected based on the assumptions presented in the literature. Namely, ring networks are considered for fronthaul/midhaul [6,7], where the number of switches does not exceed 10, as mentioned in [7]. Mesh networks are also foreseen for NGFI [6]. Eventually, reference networks DRING-16 and MESH-20 were used in C-RAN optimization studies in [16,35], respectively.

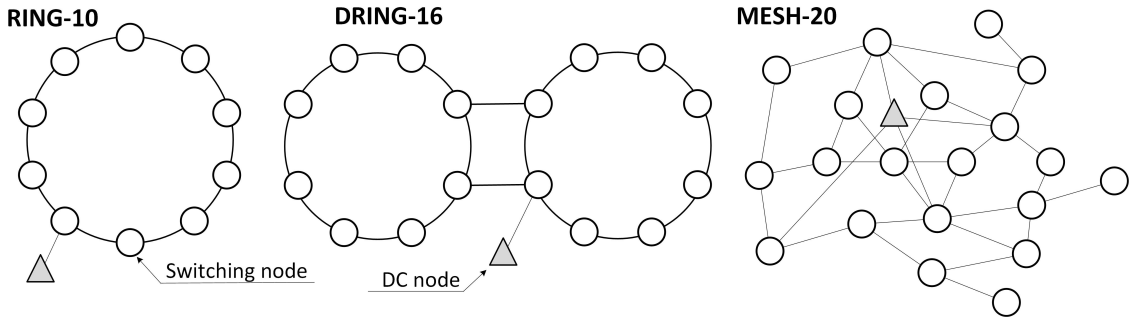

Figure 3. Network topologies: RING-10, DRING-16, and MESH-20.

Let N denote the number of switching nodes. We have $N = 10$, $N = 16$, and $N = 20$, respectively, for RING-10, DRING-16, and MESH-20. We assume that there is one PP node connected to each switching node; hence, the total number of PPs is N in the networks considered. We assume that there are R RUs, where different values of R are considered in the evaluation, and the RUs are connected to the switching nodes randomly. We assume that each RU is connected to one switch, and all the RUs connected to a given switch constitute a cluster. In Table 3, we show the values of links lengths and capacities. In particular, the length of a link is a random number generated within the limits given in Table 3. The capacities of links are in accordance to the assumptions concerning NGFI scenarios discussed in [7].

Table 3. Assumed values of link lengths and capacities.

Network Link	Link Length (km)	Link Capacity (Gbit/s)
Switch–RU	[0.2 . . . 0.5]	25
Switch–PP	[0.2 . . . 0.5]	400
Switch–DC	[10 . . . 15]	400
Switch–switch	[1 . . . 3]	100

According to [23], the total baseband processing demand of one RU of the radio system considered in Section 2.4 is about 1800 giga operations per second (GOPS). Based on the estimations presented in [23,36], in the analysis we assume that the processing loads of DU and CU are $\rho^D = 5$ and $\rho^C = 1$ processing units (PUs), respectively, where one PU represents about 300 GOPS. Due to clustering constraints, the processing capacity of a PP node (ρ) should be enough to support the total DU processing load of all RUs belonging to a cluster. Therefore, we assume that each PP node has capacity

$$\rho = \rho^D \times R^{max} \times 2 \times C, \tag{29}$$

where R^{max} denotes the size of the largest cluster of RUs in given network scenario, factor 2 is due to the DU processing of (two) associated demands (i.e., uplink and downlink) in the same PP node, and C is a PP capacity multiplier used in the analysis to scale the processing capacity of the PP node.

The routing paths between network nodes over the packet transport network have been generated using the Dijkstra shortest path algorithm. All the numerical experiments are performed on a 3.7 GHz 32-core Ryzen Threadripper-class machine with 64 GB RAM. To solve the LDCP-MILP model, we use CPLEX v.12.9 [11], which is run in a parallel mode and with default settings.

4.1. Performance of LDCP-MILP Model

We begin with evaluating the complexity of the LDCP-MILP model and the quality of obtained solutions. To this end, we solve different instances of the LDCP problem that vary in size. In particular, different network topologies, number of demands ($|\mathcal{D}|$), and PP capacities (expressed using PP capacity multiplier C) are considered. We remind that each RU involves two associated demands in the network, namely an uplink and a downlink demand; hence, the number of RUs in each scenario is $|\mathcal{D}|/2$. A 3 h computation limit is assumed in the CPLEX solver. The metrics that we report are the objective function value (z^{MILP}), computation time (T^{MILP}), and MILP optimality gap Δ^{MILP}. The optimality gap is a relative difference between z^{MILP} and the solution lower bound (z^{LB}) found in CPLEX within the computation period. Additionally, we report the obtained values of the number of active PPs ("*Active PPs*") and the overall latency of flows ("*Latency*").

In Table 4, we can see that the results obtained by solving the LDCP-MILP model are either optimal (i.e., $\Delta^{MILP} = 0\%$) or close to optimal for the majority of scenarios tested. In the cases for which near-optimal results were obtained (i.e., $\Delta^{MILP} \leq 0.06\%$), the number of active PPs is optimal (compare the most significant numbers in z^{LB} and z^{MILP}), and there is some difference in terms of latency between solution lower bounds (z^{LB}) and best objective value (z^{MILP}). We remind that the number of active PPs is the main optimization objective in the optimization problem considered. For the cases with larger optimality gaps ($\Delta^{MILP} = 3.94\%$ and $\Delta^{MILP} = 9.17\%$), we can deduce that the number of active PPs is also optimal. In particular, by subtracting the value of latency from z^{LB} and dividing the obtained number by the weighting coefficient ($A = 10^5$), we obtain a number—namely 9.6 for $\Delta^{MILP} = 3.94\%$ and 4.53 for $\Delta^{MILP} = 9.17\%$—that, rounded up, equals to the obtained value of active PPs. Note that in this analysis, rounding up is performed because the number of active PPs should be an integer number and not lower than the value resulting from the lower bound. Eventually, in the case for which $\Delta^{MILP} = 11.12\%$, the obtained number of active PPs is either optimal or it differs by not more than one from the optimal value (compare the values of z^{LB} and z^{MILP} divided by weighting coefficient A).

Table 4. Performance of the LDCP-MILP model in different network scenarios.

Scenario			Optimization Results					
Network	$\|\mathcal{D}\|$	C	z^{LB}	z^{MILP}	Δ^{MILP}	T^{MILP} (s)	**Active PPs**	**Latency [µs]**
RING-10	80	1	715,485	715,668	0.03%	10,800	7	15,668
		2	418,224	418,224	0.00%	935	4	18,224
		3	418,224	418,224	0.00%	446	4	18,224
DRING-16	64	1	973,213	1,013,174	3.94%	10,800	10	13,174
		2	513,737	514,023	0.06%	10,800	5	14,023
		3	513,514	513,514	0.00%	961	5	13,514
	80	1	1,016,681	1,017,050	0.04%	10,800	10	17,050
		2	521,205	521,443	0.05%	10,800	5	21,443
		3	*out-of-memory*					
MESH-20	40	1	805,959	805,959	0.00%	696	8	5959
		2	406,450	406,450	0.00%	295	4	6450
		3	405,822	405,822	0.00%	91	4	5822
	60	1	809,233	910,424	11.12%	10,800	9	10,424
		2	463,297	510,047	9.17%	10,800	5	10,047
		3	412,050	412,050	0.00%	348	4	12,050

In Table 4, we can also see that the complexity of solving the LDCP-MILP model decreases with increasing the available PP capacity (C). Finally, we report that when solving larger problem instances, we have encountered the problem of out-of-memory during processing of the branch-and-bound tree in CPLEX. Therefore, for solving larger problem instances, heuristic methods might be used, and we plan to develop such optimization methods in our future work.

After verifying that the quality of LDCP-MILP solutions is high, in the remainder of this work, we analyze the NGFI network using the LDCP-MILP model.

4.2. Analysis of Network Performance

In this section, we evaluate performance of the NGFI network in different network scenarios. The main performance metrics that we focus on are the number of active PP nodes and flows latencies. The evaluation is performed in the RING-10 network topology in which different numbers of RUs ($R \in \{20, 30, 40\}$) are considered. Moreover, the lengths of network links are scaled using parameter M. Namely, for link multiplier $M = 1$, the lengths of links are shown in Table 3, and the links are twice long for $M = 2$. Eventually, we scale PP capacities by considering different vales of C, where C is between 1 and 2.5.

In Figure 4, the number of active PPs as well as the average latency of the total RU-DU-CU-DC flow is shown in different RING-10 scenarios.

The network with longer links requires a higher number of active PPs than the network with basic links. This is related to larger propagation delays in the former scenario, what turns into the need for closer placement of DUs in the network with respect to the RU sites in order to meet latency requirements. Moreover, we can observe that increasing the PP capacity (parameter C) allows to activate a smaller number of PP nodes in which the DU and CU entities are placed. Note that in each network there is some value of C—for instance, $C = 1.5$ for the network with basic links and 40 RUs—above which the number of active PPs is not decreased anymore. This value can be considered as the best one since it minimizes both the number of active PP nodes and the required PP processing capacity, where both factors contribute to the network deployment cost.

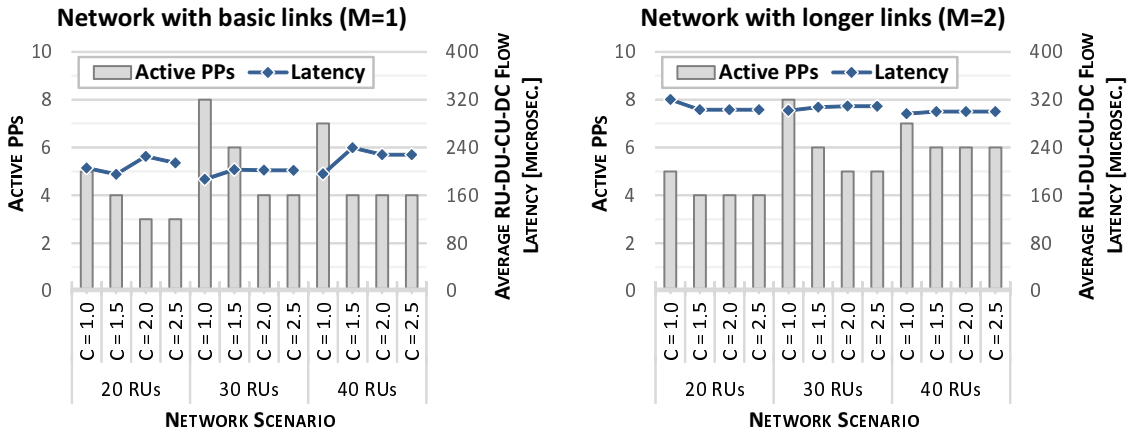

Figure 4. Number of active PPs and average overall RU-DU-CU-DC latency in network RING-10 for scenarios with basic links (**left**) and longer links (**right**).

In Figure 4, we can also see that the average latencies of the RU-DU-CU-DC flow are maintained on quite a similar level, which does not change significantly with R and C. This can be explained as following. On one hand, increasing the number of RUs in the network should increase the overall network latency. However, this effect is compensated by a larger number of active PPs and higher PP capacities (see Equation (29)) available in the scenarios with a larger number of RUs. In particular, it allows to place the DUs and CUs in less distant PP nodes and, by these means, to decrease flow latencies. To analyze this relationship in detail, in Figure 5 we present the percentage of demands for which the DU and CU processing is performed in the same PP nodes. As we can see, if a higher PP node capacity is available, which may be due to both higher number of RUs and larger C, then more demands have their DU and CU entities placed in the same PP node. Consequently, the MH flows are not present in the network, and they do not contribute to the overall network latency. Indeed, as shown in Figure 6, the maximum latencies of midhaul flows are equal to 0 in the scenarios in which joint DU/CU processing is performed for all demands (i.e., for the scenarios reaching 100% of joint processing in Figure 5).

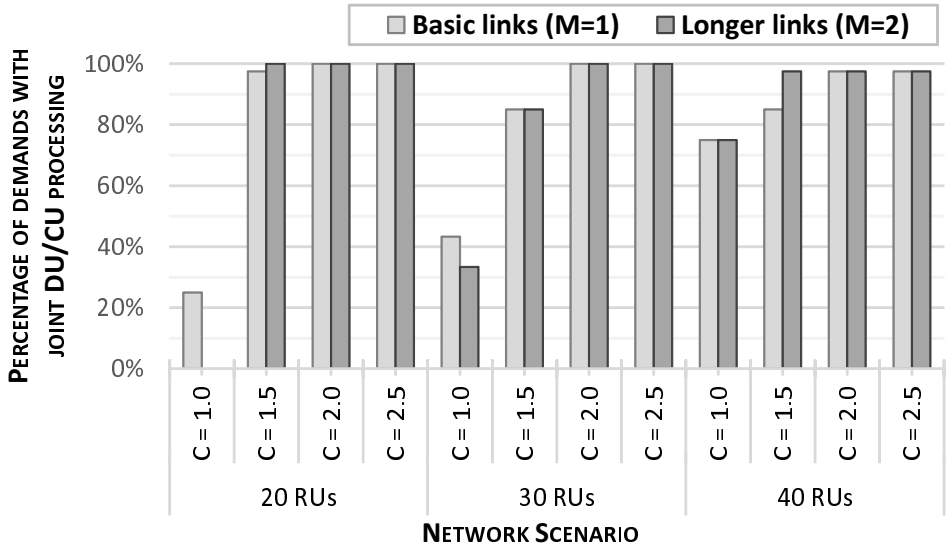

Figure 5. Percentage of demands with DU and CU processing performed in the same PP node in network RING-10.

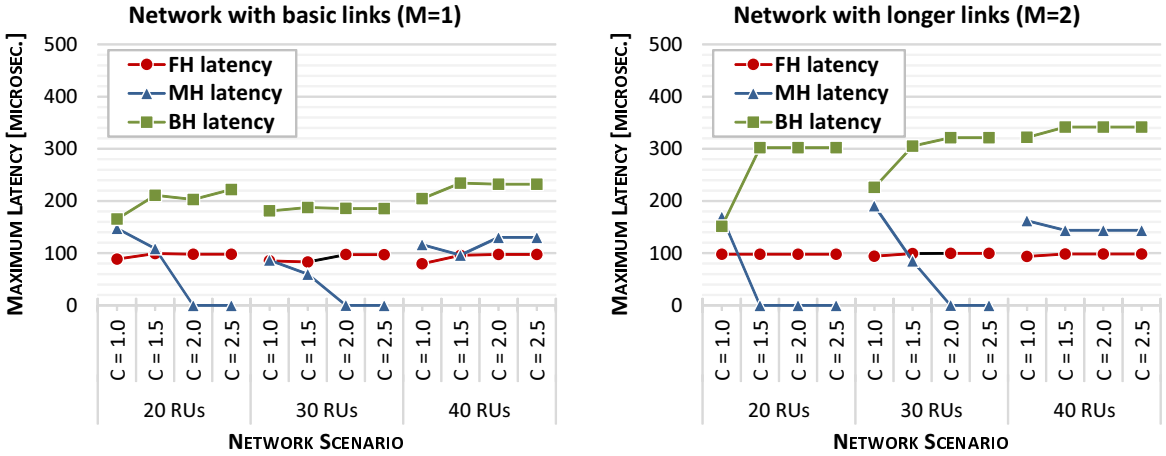

Figure 6. Maximum latencies of fronthaul, midhaul, and backhaul flows in network RING-10 for scenarios with basic links (**left**) and longer links (**right**).

The maximum latencies of FH flows shown in Figure 6 are maintained below 100 µs, which is the allowable limit in FH. This validates the correct implementation of latency constraints in the MILP model. Moreover, the MH and BH latencies are below 190 µs and 360 µs, respectively, which is far below the latency limits assumed for these flows (1 ms and 2 ms, respectively), even in the largest network (40 RUs) with longer links. Lower values of maximum BH flow latency for 30 RUs and $C \geq 1.5$ in the basic scenario ($M = 1$), when compared to the scenario with 20 RUs, might be due to a particular placement of CU entities with respect to the DC node. Namely, if the most distant CU entity is placed closer to the DC node, then this results in a lower maximum propagation delay of a BH flow in the former scenario than in the latter scenario. Note that for 30 RUs, we have a higher number of active PPs in the network (see Figure 4), which increases the chance to place the CUs closer to the DC. Finally, the MH and BH latencies are higher in the scenario with longer links ($M = 2$), which is due to larger propagation delays.

4.3. Evaluation of Larger Network Topologies

To complete the analysis, we evaluate two larger networks: DRING-16 and MESH-20. The number of randomly located RUs is 40 and 30, and the largest cluster of RUs consists of 5 and 4 RUs, respectively, in DRING-16 and MESH-20. As in the RING-10 network, in the analysis we considered different values of the PP capacity multiplier, where C is between 1 and 3.

In Figure 7, we show maximum latencies of fronthaul, midhaul, and backhaul flows. Moreover, we report the number of active processing pools in network DRING-16 and in network MESH-20. We can see that the number of active PP nodes decreases if the PP capacity increases. The maximum FH latencies are below 100 µs, which indicates that the solutions obtained are correct. In both networks, the maximum latencies of MH and BH flows are below 140 µs and 210 µs, respectively, which is much lower than the allowable limits. Slightly lower numbers of active PPs in MESH-20 may be explained by higher connectivity of nodes in the MESH-20 network when compared to the DRING-16 network.

In Figure 8, we present the overall capacity of active PP nodes, expressed in terms of processing units (PUs), in networks DRING-16 and MESH-20. Again, we provide the number of active PPs in the figure. In both networks, we can see that there is some value of C for which both the number of active PPs is minimized and the overall capacity of the PP nodes is either the lowest (for $C = 2$ in DRING-16) or near to the lowest value (for $C = 2.5$ in MESH-20). This value of C can be considered as the best one since it minimizes the deployment cost of PPs in the network.

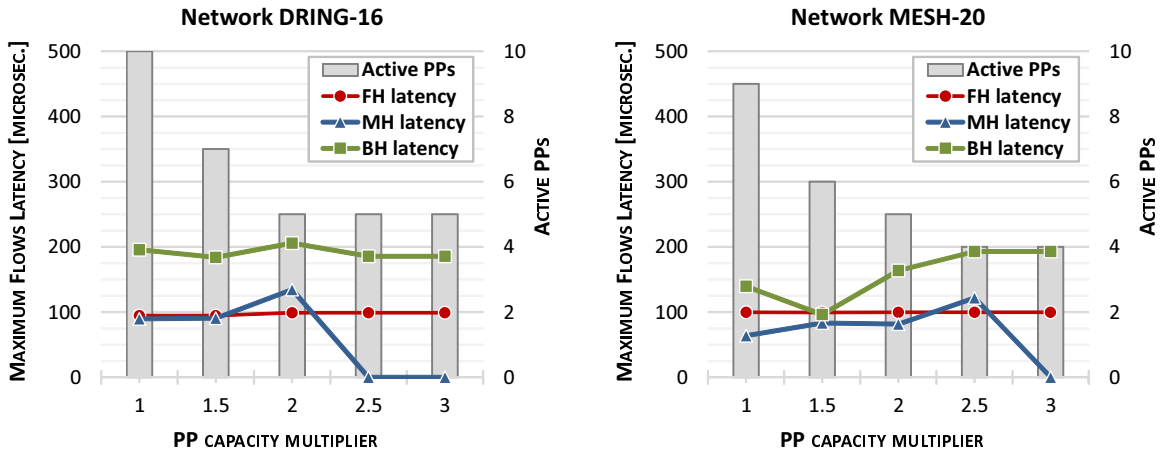

Figure 7. Maximum latencies of fronthaul, midhaul, and backhaul flows, and number of active PPs for different values of PP capacity multiplier (C) in networks DRING-16 (**left**) and MESH-20 (**right**).

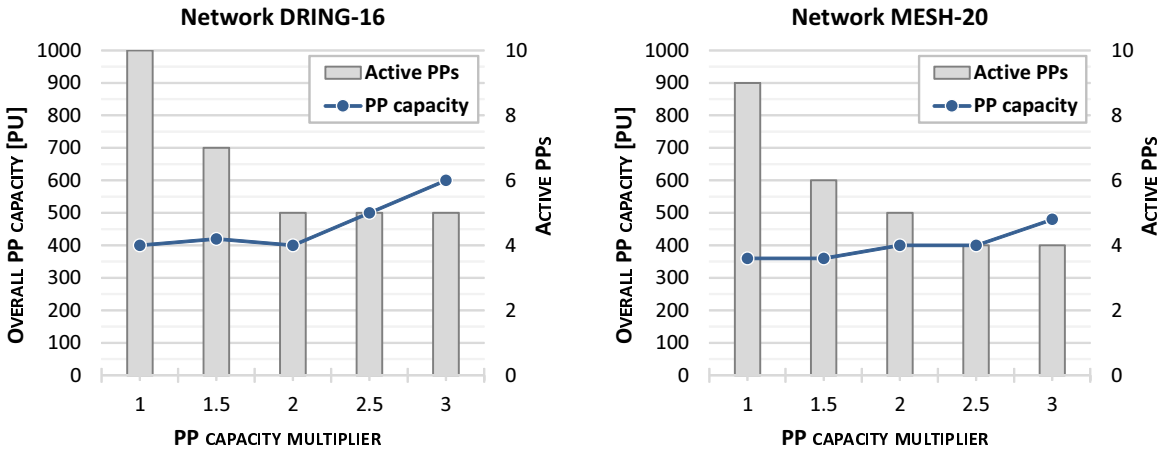

Figure 8. Overall capacity of active PP nodes and number of active PPs for different values of PP capacity multiplier (C) in networks DRING-16 (**left**) and MESH-20 (**right**).

Similarly as in the RING-10 network, in Figure 7 we can see that maximum MH flow latencies are equal to 0 if the PP capacity is large enough, namely, for $C = 2.5$ in DRING-16 and for $C = 3$ in MESH-20. As shown in Figure 9a, these cases correspond to the scenarios in which 100% of demands have their DU and CU processing performed in the same PP node and, consequently, the MH flows are not present in the network. In Figure 9a, we can also see that with the smallest required PP capacity (i.e., for $C = 1$), about 40% of demands have their baseband processing performed in the same PP node. Increasing the PP capacity by 50% (i.e., for $C = 1.5$), the percentage of joint DU/CU processing increases to 80% of demands. Finally, in Figure 9b, we analyze the average usage of available PP capacity of active PP nodes in different network scenarios. In general, we can observe that the average percentage usage of PP capacity tends to decrease with C. This relationship can be explained by the fact that higher values of C lead to the increase of the overall processing capacity in the network if the number of active PPs does not decrease significantly (as for $C > 2$ in both networks). Since the DU/CU processing demand in a given network scenario is fixed and the overall processing capacity increases, then the average PP usage must decrease.

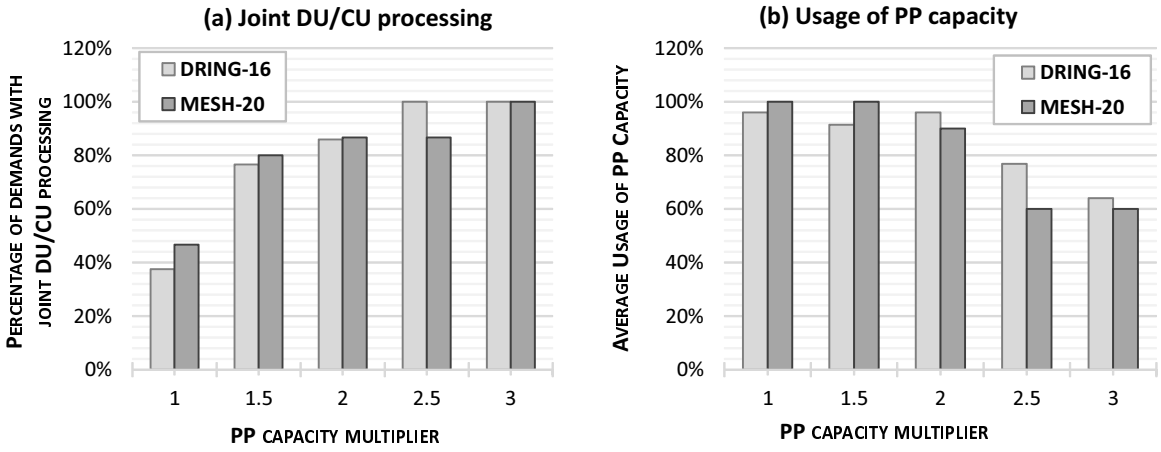

Figure 9. (**a**) Percentage of demands with DU and CU processing performed in the same PP node (left) and (**b**) average usage of PP capacity of active PP nodes (right) in a function of PP capacity multiplier (C) in networks DRING-16 and MESH-20.

5. Conclusions

We have focused on latency-aware DU and CU placement (LDCP) in packet-based NGFI networks. The LDCP optimization problem was modeled as a mixed-integer linear programming problem. We have made use of the latency model that estimates worst-case latencies of flows to guarantee that the traffic flows carried over the packet-switched network satisfy latency constraints. To evaluate performance of the LDCP-MILP model and NGFI network, we have considered different network scenarios varying in topologies, the number of RUs, PP capacities, and link lengths were considered.

Solving the LDCP-MILP model is feasible for network instances consisting of some tens of RUs and about 20 switching nodes. Indeed, for the network scenarios considered in this work, we have obtained good-quality solutions when solving the model. At the same time, for larger problem instances we have encountered the issue of lack of memory during solving the model by a commercial mixed-integer programming solver (CPLEX). Note that introduction of additional constraints into the MILP model, such as routing constraints, will make the problem more complex. Therefore, optimization of larger networks and extended network scenarios will require efficient heuristics and/or the application of advanced MILP optimization techniques, such as column generation and cut generation. In our previous works, we have shown the effectiveness of hybrid optimization algorithms—combining different processing and optimization techniques—in solving large problem instances that are difficult to be treated by mathematical integer programming solvers [37]. We plan to develop such optimization methods in our future work in the context of packet-based NGFI networks.

As we have shown in the analysis of network performance, a higher number of active PP nodes is required in larger networks to keep flow latencies within allowable limits. At the same time, the number of active PPs decreases if more capacity is available in the PP nodes. In general, the latency of the overall (RU-DU-CU-DC) flow does not change significantly if more PP capacity is available. It comes from two opposite effects that compensate each other, namely the decrease of the number of active PP nodes and the increase in joint DU/CU processing. In particular, the former effect may increase the BH flow latencies, while the latter decreases the total latency of MH flows. Eventually, we have observed that proper selection of the PP node capacity may lead to minimization of the number of active PP nodes without a significant overhead in the total PP capacity deployed in the network. This in turn results in minimization of the network cost.

In future works, we will focus on different problems that exist in NGFI networks and that require dedicated optimization algorithms, including network slicing or network survivability. We will also consider more diversified network scenarios, including the networks in which some of the links are wireless links. Finally, we plan to work on improving MILP formulations and to make use of

advanced optimization methods, with the aim to develop optimization methods applicable to larger network scenarios.

Funding: This research was funded by National Science Centre, Poland, under grant number 2018/31/B/ST7/03456.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Agiwal, M.; Roy, A.; Saxena, N. Next Generation 5G Wireless Networks: A Comprehensive Survey. *IEEE Commun. Surv. Tutorials* **2016**, *18*, 1617–1655. [CrossRef]
2. The 3rd Generation Partnership Project (3GPP). Available online: http://www.3gpp.org/ (accessed on 28 September 2020).
3. Peng, M.; Sun, Y.; Li, X.; Mao, Z.; Wang, C. Recent Advances in Cloud Radio Access Networks: System Architectures, Key Techniques, and Open Issues. *IEEE Commun. Surv. Tutorials* **2016**, *18*, 2282–2308. [CrossRef]
4. Alimi, I.A.; Teixeira, A.; Monteiro, P. Towards an Efficient C-RAN Optical Fronthaul for the Future Networks: A Tutorial on Technologies, Requirements, Challenges, and Solutions. *IEEE Commun. Surv. Tutor.* **2018**, *20*, 708–769. [CrossRef]
5. Gomes, N.J.; Sehier, P.; Thomas, H.; Chanclou, P.; Li, B.; Munch, D.; Jungnickel, V. Boosting 5G through Ethernet. *IEEE Vehic. Technol. Mag.* **2018**, *55*, 74–84. [CrossRef]
6. *ITU-T Technical Report*; Transport Network Support of IMT-2020/5G; International Telecommunication Union: Geneva, Switzerland, 2018.
7. IEEE. IEEE Standard for Packet-Based Fronthaul Transport Networks. Available online: https://standards.ieee.org/project/1914_1.html (accessed on 28 September 2020).
8. IEEE. *802.1CM-2018—IEEE Standard for Local and Metropolitan Area Networks—Time-Sensitive Networking for Fronthaul*; IEEE: Middlesex, NJ, USA, 2018.
9. Garcia-Saavedra, A.; Salvat, J.X.; Li, X.; Costa-Perez, X. WizHaul: On the Centralization Degree of Cloud RAN Next Generation Fronthaul. *IEEE Trans. Mob. Comput.* **2018**, *17*, 2452–2466. [CrossRef]
10. O-RAN Alliance. Available online: https://www.o-ran.org/ (accessed on 28 September 2020).
11. IBM. CPLEX Optimizer. Available online: http://www.ibm.com/ (accessed on 28 September 2020).
12. Carapellese, N.; Tornatore, M.; Pattavina, A.; Gosselin, S. BBU Placement over a WDM Aggregation Network Considering OTN and Overlay Fronthaul Transport. In Proceedings of the 2015 European Conference on Optical Communication (ECOC), Valencia, Spain, 27 September–1 October 2015.
13. Musumeci, F.; Bellanzon, C.; Tornatore, N.C.M.; Pattavina, A.; Gosselin, S. Optimal BBU Placement for 5G C-RAN Deployment over WDM Aggregation Networks. *IEEE J. Lightw. Technol.* **2016**, *34*, 1963–1970. [CrossRef]
14. Velasco, L.; Castro, A.; Asensio, A.; Ruiz, M.; Liu, G.; Qin, C.; Yoo, S.B. Meeting the Requirements to Deploy Cloud RAN Over Optical Networks. *OSA/IEEE J. Opt. Commun. Netw.* **2017**, *9*, B22–B32. [CrossRef]
15. Wong, E.; Grigoreva, E.; Wosinska, L.; Machuca, C.M. Enhancing the Survivability and Power Savings of 5G Transport Networks based on DWDM Rings. *OSA/IEEE J. Opt. Commun. Netw.* **2017**, *9*, D74–D85. [CrossRef]
16. Khorsandi, B.M.; Raffaelli, C. BBU location algorithms for survivable 5G C-RAN over WDM. *Comput. Netw.* **2018**, *144*, 53–63. [CrossRef]
17. Liu, J.; Zhou, S.; Gong, J.; Niu, Z.; Xu, S. Graph-based Framework for Flexible Baseband Function Splitting and Placement in C-RAN. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 8–12 June 2015.
18. Koutsopoulos, I. Optimal Functional Split Selection and Scheduling Policies in 5G Radio Access Networks. In Proceedings of the 2017 IEEE International Conference on Communications Workshops (ICC Workshops), Paris, France, 21–25 May 2017.
19. Ejaz, W.; Sharma, S.K.; Saadat, S.; Naeem, M.; Anpalagan, A.; Chughtai, N.A. A Comprehensive survey on Resource Allocation for CRAN in 5G and Beyond Networks. *J. Net. Comput. Appl.* **2020**, *160*, 1–24. [CrossRef]

20. Wang, X.; Alabbasi, A.; Cavdar, C. Interplay of Energy and Bandwidth Consumption in CRAN with Optimal Function Split. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017.
21. Alabbasi, A.; Wang, X.; Cavdar, C. Optimal Processing Allocation to Minimize Energy and Bandwidth Consumption in Hybrid CRAN. *IEEE Trans. Green Commun. Netw.* **2018**, *2*, 545–555. [CrossRef]
22. Yu, H.; Musumeci, F.; Zhang, J.; Xiao, Y.; Tornatore, M.; Ji, Y. DU/CU Placement for C-RAN over Optical Metro-Aggregation Networks. In Proceedings of the 23rd Conference on Optical Network Design and Modelling, Athens, Greece, 13–16 May 2019.
23. Xiao, Y.; Zhang, J.; Ji, Y. Can Fine-grained Functional Split Benefit to the Converged Optical-Wireless Access Networks in 5G and Beyond? *IEEE Trans. Netw. Serv. Manag.* **2020**, *17*, 1774–1787. [CrossRef]
24. Nakayama, Y.; Hisano, D.; Kubo, T.; Fukada, Y.; Terada, J.; Otaka, A. Low-Latency Routing Scheme for a Fronthaul Bridged Network. *OSA/IEEE J. Opt. Commun. Netw.* **2018**, *10*, 14–23. [CrossRef]
25. Hisano, D.; Nakayama, Y.; Kubo, T.; Uzawa, H.; Fukada, Y.; Terada, J. Decoupling of Uplink User and HARQ Response Signals to Relax the Latency Requirement for Bridged Fronthaul Networks. *OSA/IEEE J. Opt. Commun. Netw.* **2019**, *11*, B26–B36. [CrossRef]
26. Klinkowski, M.; Mrozinski, D. Latency-Aware Flow Allocation in 5G NGFI Networks. In Proceedings of the 2020 22nd International Conference on Transparent Optical Networks (ICTON), Bari, Italy, 19–23 July 2020.
27. Klinkowski, M. Optimization of Latency-Aware Flow Allocation in NGFI Networks. *Comp. Commun.* **2020**, *161*, 344–359. [CrossRef]
28. 3GPP. *Study on New Radio Access Technology: Radio Access Architecture and Interfaces*; Technical Report 38.801, v14.0.0; European Telecommunications Standards Institute: Sophia Antipolis, France, 2017.
29. Anritsu. 1914.3 (RoE) eCPRI Transport White Paper. 2018. Available online: https://dl.cdn-anritsu.com/en-en/test-measurement/files/Technical-Notes/White-Paper/mt1000a-ecpri-er1100.pdf (accessed on 28 September 2020).
30. Imran, M.A.; Zaidi, S.A.R.; Shakir, M.Z. *Access, Fronthaul and Backhaul Networks for 5G & Beyond*; Institution of Engineering and Technology: London, UK, 2017.
31. IEEE. *802.1CM-2018—IEEE Standard for Local and Metropolitan Area Networks—Bridges and Bridged Networks*; IEEE: Middlesex, NJ, USA, 2018.
32. Perez, G.O.; Larrabeiti, D.; Hernandez, J.A. 5G New Radio Fronthaul Network Design for eCPRI-IEEE 802.1CM and Extreme Latency Percentiles. *IEEE Access* **2019**, *7*, 82218–82229. [CrossRef]
33. IEEE 1914 Working Group. Fronthaul Dimensioning Tool. Available online: https://sagroups.ieee.org/1914/p1914-1/ (accessed on 28 September 2020).
34. Garey, M.R.; Johnson, D.R. *Computers and Intractability: A Guide to the Theory of NPCompleteness*; W H Freeman & Co: New York, NY, USA, 1979.
35. Khorsandi, B.M.; Tonini, F.; Raffaelli, C. Centralized vs. Distributed Algorithms for Resilient 5G Access Networks. *Phot. Netw. Commun.* **2019**, *37*, 376–387. [CrossRef]
36. Shehata, M.; Elbanna, A.; Musumeci, F.; Tornatore, M. Multiplexing Gain and Processing Savings of 5G Radio-Access-Network Functional Splits. *IEEE Trans. Green Commun. Netw.* **2018**, *2*, 982–991. [CrossRef]
37. Klinkowski, M.; Walkowiak, K. An Efficient Optimization Framework for Solving RSSA Problems in Spectrally and Spatially Flexible Optical Networks. *IEEE/ACM Trans. Netw.* **2019**, *27*, 1474–1486. [CrossRef]