Optimal QoS-Aware Allocation of Virtual Network Resources to Mixed Mobile-Optical Network Slices

Citation for the original published paper (version of record):
Keshavarz, M., Hadi, M., Lashgari, M. et al (2021). Optimal QoS-Aware Allocation of Virtual Network Resources to Mixed Mobile-Optical Network Slices. Proceedings - IEEE Global Communications Conference, GLOBECOM.
http://dx.doi.org/10.1109/GLOBECOM46510.2021.9685414

(article starts on next page)

Optimal QoS-Aware Allocation of Virtual Network Resources to Mixed Mobile-Optical Network Slices

Mohammad Hossein Keshavarz, Mohammad Hadi, Maryam Lashgari, Mohammad Reza Pakravan, and Paolo Monti

***Abstract*—Slicing allows 5G networks to accommodate services with different needs over a unified physical network infrastructure. Particularly, in radio access networks (RANs), the baseband processing functions of different slices are treated as virtual applications running on shared compute nodes. Managing this cloud-native network in a cost-effective way requires tightly coupled control of connectivity and processing resources. This paper proposes an optimization problem to minimize the overall cost while guaranteeing distinct latency and reliability requirements of different network slices deployed over the infrastructure. This is achieved by choosing for each service, the functional split and transmission parameters that best adapt to the connectivity and processing resources available in the infrastructure. Results demonstrate that the proposed flexible resource allocation scheme provides considerable cost saving (up to 2.4 times) and reduced request blocking (up to 2 times) compared to conventional fixed deployment techniques. The reliability requirements can be guaranteed by packet duplication and/or virtualized forward error correction at the expense of consuming connectivity and/or processing resources, respectively. The flexible assurance of the reliability requirements in the proposed scheme contributes considerably to the achieved improvements on cost saving and request blocking.**

***Keywords*—Network Function Virtualization, Network Slicing, Functional split, Optical transport networks, Reliability, 5G**

I. Introduction

The continuous growth of mobile network traffic forces operators to use smaller cells which require installing more but statically less utilized equipment. To overcome the growing costs imposed by the increasing number of underutilized deployed small cells, centralized radio access network (C-RAN) was used in 4G. In C-RAN, traditional base stations are disaggregated into remote radio unit (RU) and baseband unit (BBU). Remote RUs interface with antennas and are responsible for radio frequency (RF) operations and analog/digital conversion. BBUs are deployed in processing pools (PPs) of a central unit (CU) and perform digital signal processing and protocol functions. The resource pooling allows more efficient usage of installed resources and leads to enhanced performance using coordination between neighboring cells [1].

The huge offered bit rate of fifth generation (5G) requires more bandwidth on the link between the remote RU and BBU pool referred to as fronthaul link. Moreover, diverse use cases of 5G deployment scenarios impose different requirements on the radio access network (RAN). These issues along with the ubiquitous use of network function virtualization in 5G paved the way for the advent of functional splits in RANs. In functional splits, the baseband processing is divided into several virtual network functions (VNFs) and the location of processing each VNF is decided according to the network state and user requirements. Flexible functional split selection provides an applicable trade off between the imposed requirements on the fronthaul, and complexity and utilization of the deployed resources. Functional split selection makes resource allocation in RAN more intricate [2], [3].

Two other trends make cost-efficient resource allocation in 5G RAN even more complex. First, to provide flexibility in serving diverse applications of 5G, network slicing was introduced. Network slicing enables the possibility of using different VNFs, configurations, and/or resource allocations for each slice. Second, 5G is supposed to serve delay-sensitive applications as close to the user as possible inside the RAN [4]. Mobile edge computing (MEC) is a viable solution to facilitate delivery of services with strict latency requirements.

Many works have addressed cost-efficient allocation of connectivity and processing resources in RANs. The authors in [2] solved the resource allocation problem in a wavelength devision multiplexing (WDM) network with the freedom to distribute baseband processing functions (BPFs) for each flow in multiple locations, and analyzed how this fine-grained functional splitting is beneficial. The work in [5] addressed the resource allocation problem in a setting with multiple CUs. The authors in [6] investigated the functional split selection problem on a per slice basis taking different latency requirements between slices into account. In [7], MEC was considered as an additional BPF and the combined problem of MEC placement and functional split selection was tackled.

Operators eagerly intend to use optical transport network (OTN) for the fronthaul part of the RAN [8]. However, neither of the surveyed works have considered a detailed model of the optical network. Further, most of the papers on the RAN resource allocation focus on the latency requirements of various network slices, especially ultra-reliable low latency communication (URLLC), without any attention to service reliability requirements, as an important QoS metric [9]. In this domain, [10] considered the service reliability requirements and compared dedicated and shared path protection for URLLC services. Traditional path protection methods are mostly intended to only guarantee availability. However, the deployed availability resources can be employed to decrease packet error rate (PER) by a simple processing of the data received on both working and protection paths. Although this efficient feature can be specially useful for URLLC services with stringent PER limits, a few rare research works have exploited it. More recently, the authors in [11] considered end-to-end packet duplication to increase reliability performance of URLLC service. They investigated packet duplication with either wavelength isolation or link isolation and compared their corresponding service request blocking ratio. Other works that address reliability requirements of URLLC service mostly concentrate on the application layer and exploit redundancy and/or coding [12].

M. H. Keshavarz, M. Hadi, and M. R. Pakravan are with the Department of Electrical Engineering, Sharif University of Technology, Tehran, Iran. P. Monti and M. Lashgari are with the Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden. This work was in part supported by VINNOVA as part of the project "Smart city concepts in Curitiba - Low-carbon transport and mobility in a digital society" (2019-04893) and by the MSCA-ITN project "5G STEP FWD" with funding from the European Union's Horizon 2020 research under grant agreement number 722429.

Historically, optical networks provided robust ways of guaranteeing reliability and availability, which became more controllable with the appearance of advanced forward error corrections (FECs) in OTN standard [13]. Such FEC schemes provide an acceptable coding gain within a considerably low processing delay. Therefore, if OTN serves as the fronthaul network, adjustable FEC parameters can be properly tuned to simultaneously guarantee latency and PER requirements for a reasonable amount of resource usage and processing cost.

In this work, we propose a QoS-aware resource assignment scheme for joint allocation of connectivity and processing resources in a RAN built on an OTN with adjustable FEC parameters. As an important feature, the FEC processing and packet duplication are incorporated into the scheme to allow cost-efficient assurance of reliability requirements. We assume that FEC processing, as a main source of operating expense in the fronthaul [14], [15], can be performed on the same processing pool used for other BPFs. As far as we know, this way of FEC processing leads to an emerging level of virtualization in optical networks as well as in integration of optical and radio access networks, which is not considered in conventional rigid optical networks. The proposed scheme contributes to the ongoing researches as follows.

- The scheme concurrently determines functional splits, network configuration, and MEC allocation to cost-efficiently guarantee QoS requirements of all network slices.
- The scheme is established by taking a detailed description of the underlying optical network, reliability measures, and packet duplication into account.
- The performance of the scheme is evaluated compared to different benchmark counterparts.
- The scheme provides a constructive synergy between optical and radio domain and shows how this synergy leads to cost-efficient resource allocation.

The rest of the paper is organized as follows. The system model is described in Section II while the resource allocation scheme is mathematically formulated in Section III. The performance of the scheme is validated in Section IV. Finally, in Section V, the paper ends after a concise conclusion.

II. Problem Formulation

A. Variable Constraints

These equations describe conditions which define each of optimization variables:

$$\sum_{t} x_{i,i',s,t} = 1 - b_{i,i',s} \quad \forall i, i', s \tag{1}$$

$$x_{i,i',s,1} \leqslant M'_{i,i',s} \forall i, i', s \tag{2}$$

$$\sum_{t=8,9,10} x_{i,i',s,t} \leqslant M_{i,i',s} \quad \forall i, i', s \tag{3}$$

$$\sum_{m} f_{i,k,l,m} = 1 \quad \forall i, k, l \tag{4}$$

$$f_{i,0,l,m} = 1 \quad \forall i, l, m \tag{5}$$

$$\sum_{k,l} y_{i,i',s,k,l} = 1 - b_{i,i',s} - \sum_{t=1,8,9,10} x_{i,i',s,t} \quad \forall i, i', s \tag{6}$$

$$\sum_{k',l} y'_{i,i',s,k',l} = 1 - b_{i,i',s} - \sum_{t=1,8,9,10} x_{i,i',s,t} \quad \forall i, i', s \tag{7}$$

$$y_{i,i',s,k,l} \leqslant \sum_{m \geqslant 1} f_{i,k,l,m} \quad \forall i, i', s, k, l \tag{8}$$

$$y'_{i,i',s,k',l} \leqslant \sum_{m \geqslant 1} f_{i,k',l,m} \quad \forall i, i', s, k', l \tag{9}$$

$$\sum_{l} \left[y_{i,i',s,k,l} + y'_{i,i',s,k,l} \right] \leqslant 1 \quad \forall i, i', s, k \tag{10}$$

Equation 1 forces choosing one of the ten functional splits for each slice in each RU. We may choose not to serve an slice based on our other constraints(capacity or QoS constraints) hence we added the variable $b_{i,i',s}$ to indicate this. Equations 2 and 3 make sure that we use MEC in RU or DU only if processing application logic for the respective slice is possible in that nodes. Next five equations fix transmission parameters for each slice. First 4 chooses whether to use a lightpath and fixes modulation level and FEC for it. Setting $f_{i,k,l,0}$ to 1 is reserved for not using a wavelength on a link in the node. Equation 5 is used as convenience for simplifying later equations we reserved $k = 0$ for not using a second path and this sets the value for f in that case to 1. Then 6 and 7 assign primary and secondary (in case of PD) lightpaths for each slice if we choose to serve that slice and if we do not use MEC for that slice. Note that summation bounds on k' are different from k, we used $k' = 0$ as a proxy for not using a secondary path here. 8 and 9 are added here to facilitate reading (and in simulations to accelerate solving) but their constraint is satisfied by 14. These equations limit chosen lightpath for an slice to lightpaths which are active in the node as specified by f. At last 10 forces the optimization to use different paths for primary and secondary flows of an slice.

B. QoS constraints

$$y_{i,i',s,k,l} + x_{i,i',s,t} + f_{i,k,l,m} \leqslant 2 \quad \forall i, i', s, k, l, m, t : T_{i,i',s,t} \leqslant D_{i,k,m} \tag{11}$$

$$y'_{i,i',s,k',l} + x_{i,i',s,t} + f_{i,k',l,m} \leqslant 2 \quad \forall i, i', s, k', l, m, t : T_{i,i',s,t} \leqslant D_{i,k',m} \tag{12}$$

$$\sum_{l} \left(y_{i,i',s,k,l} + y'_{i,i',s,k',l} + f_{i,k,l,m} + f_{i,k',l,m} \right) \leqslant 3 \quad \forall i, i', s, k, k', m, m' \quad \text{where} \quad H_{i,i',s} < R_{i,k,k',m} \tag{13}$$

The first two Equations limit using paths with larger delay than slice's tolerable delay for each slice. They limit choosing combinations of y, x, f where the combination leads to excessive delay. Equation 13, does the same job for reliability requirements of each slice.

C. Capacity constraints

$$\sum_{i',s,t} \left[z_{i,i',s,k,l,t} g_{i,i',s,t} + z'_{i,i',s,k',l,t} g_{i,i',s,t} \right] \leqslant \sum_{m} f_{i,k,l,m} A_m \quad \forall i, k, l \tag{14}$$

$$\sum_{i,k,m \geqslant 1} f_{i,k,l,m} U_{i,k,j} \leqslant 1 \quad \forall j, l \tag{15}$$

TABLE I: List of indices, parameters, and variables along with their corresponding names. $\mathbb{N}$, $\mathbb{W}$, and $\mathbb{R}$ are sets of natural, whole, and real numbers, respectively. $\mathbb{X}^c$ means all numbers in set $\mathbb{X}$ that satisfies condition c.

Type	Notation	Name
Indices	$i \in \mathbb{N}^{\leqslant N}$	DU Node index
	$i' \in \mathbb{N}^{\leqslant n_i}$	RU Node index
	$j \in \mathbb{N}^{\leqslant F}$	Link index
	$s \in \mathbb{N}^{\leqslant S_{i,i'}}$	Slice index
	$t \in \mathbb{W}^{\leqslant 10}$	Split index
	$l \in \mathbb{N}^{\leqslant L}$	Wavelength index
	$k \in \mathbb{N}^{\leqslant P_i}$	Path index
	$k' \in \mathbb{W}^{\leqslant P_i}$	secondary Path index, zero for nothing
	$m \in \mathbb{W}^{\leqslant 6}$	Transmission mode index
Parameters	$P_i \in \mathbb{N}$	Number of paths between DU i and CU
	$N \in \mathbb{N}$	Number of DU nodes
	$F \in \mathbb{N}$	Number of fiber links
	$L \in \mathbb{N}$	Number of wavelengths
	$S_{i,i'} \in \mathbb{N}^{\leqslant S}$	Number of slices in RU i' of DU i
	$C'_{i,i'} \in \mathbb{R}^{\geqslant 0}$	Available processing capacity in RU
	$C_i \in \mathbb{R}^{\geqslant 0}$	Available processing capacity in DU
	$C_0 \in \mathbb{R}^{\geqslant 0}$	Available processing capacity in CU
	$g_{i,i',s,t} \in \mathbb{R}$	required data rate between DU and CU
	$A_m \in \mathbb{R}$	Data rate of lightpath with $f_{i,k,l,m}$
	$E \in \mathbb{R}^{\geqslant 0}$	Overall cost
	$E^B \in \mathbb{R}^{\geqslant 0}$	Total BPF processing cost
	$E^F \in \mathbb{R}^{\geqslant 0}$	Total FEC processing cost
	$E^N \in \mathbb{R}^{\geqslant 0}$	Total connectivity cost
	$\alpha \in \mathbb{R}^{\geqslant 0}$	Equivalent processing cost in RUs
	$\beta \in \mathbb{R}^{\geqslant 0}$	Equivalent processing cost in DUs
	$\gamma \in \mathbb{R}^{\geqslant 0}$	Equivalent connectivity cost
	$M'_{i,i',s} \in \mathbb{R}^{\geqslant 0}$	MEC possibility in RU
	$M_{i,i',s} \in \mathbb{R}^{\geqslant 0}$	MEC possibility in DU
	$U_{i,k,j} \in \{0,1\}$	Link path membership indicator
	$D_{i,k,m} \in \mathbb{R}^{\geqslant 0}$	Achievable delay
	$R_{i,k,k',m,m'} \in (0,1]$	Achievable PER
	$T_{i,i',s,t} \in \mathbb{R}^{\geqslant 0}$	Tolerable delay
	$H_{i,i',s} \in (0,1]$	Tolerable PER
	$O'_{i,i',s,t} \in \mathbb{R}^{\geqslant 0}$	Required BPF processing capacity at RU
	$O_{i,i',s,t} \in \mathbb{R}^{\geqslant 0}$	Required BPF processing capacity at DU
	$Q_{i,i',s,t} \in \mathbb{R}^{\geqslant 0}$	Required BPF processing capacity at CU
	$V_{i,i',s,m} \in \mathbb{R}^{\geqslant 0}$	Required FEC processing capacity
Variables	$x_{i,i',s,t} \in \{0,1\}$	Functional split selector
	$b_{i,i',s} \in \{0,1\}$	Blocking selector
	$y_{i,i',s,k,l} \in \{0,1\}$	Primary lightpath selector
	$y'_{i,i',s,k',l} \in \{0,1\}$	Secondary lightpath selector
	$f_{i,k,l,m} \in \{0,1\}$	Transmission mode selector
	$z_{i,i',s,k,l,t} \in \{0,1\}$	(aux.) $y_{i,i',s,k,l}x_{i,i',s,t}$
	$z'_{i,i',s,k',l,t} \in \{0,1\}$	(aux.) $y'_{i,i',s,k',l}x_{i,i',s,t}$

Each lightpath must support data rate of all slices it carry as described in 14 and each wavelength should be used only in one active lightpath as described in 15.

The next three equations set appropriate capacity limits on the amount of processing performed in each RU, DU and the CU.

$$\sum_{s,t} x_{i,i',s,t} O'_{i,i',s,t} \leqslant C'_{i,i'} \quad \forall i, i' \tag{16}$$

$$\sum_{i',s,t} x_{i,i',s,t} O_{i,i',s,t} \quad + \sum_{k,m} f_{i,k,l,m} V_m \leqslant C_i \quad \forall i \tag{17}$$

$$\sum_{i,i,s,t} x_{i,i',s,t} Q_{i,i',s,t} + \sum_{i,k,l,m} f_{i,k,l,m} V_m \leqslant C_0 \tag{18}$$

D. Objective Function

Equation

$$\min_{\substack{x_{i,i',s,t},\, f_{i,k,l,m},\\ y_{i,i',s,k,l},\, y'_{i,i',s,k',l}, z_{i,i',s,t,m}}} \quad E^B + E^F + E^N \tag{19}$$

Describes optimization goal. In fact we want to minimize overall cost of the network. This cost is comprised of the cost of baseband processing, FEC processing and transmission which are expressed below in equations 20,26b and 22 respectively.

$$E^B = \sum_{i,i,s,t} \big[\alpha x_{i,i',s,t} O'_{i,i',s,t} + \beta x_{i,i',s,t} O_{i,i',s,t} + x_{i,i',s,t} Q_{i,i',s,t}\big] \tag{20}$$

$$E^F = (\beta + 1) \sum_{i,k,l,m} f_{i,k,l,m} V_m \tag{21}$$

$$E^N = \gamma \sum_{i,k,l,m \geqslant 1} f_{i,k,l,m} \tag{22}$$

III. Problem Formulation

We aim to allocate connectivity and processing resources such that the overall network cost is minimized while physical restrictions, capacity limitations, and QoS requirements of slices are satisfied. This problem can be formulated as an integer linear program (ILP) whose variables, constraints, and objective function are described in the next sub-sections. The variables, parameters, and indices of the formulation are summarized in Tab. II.

A. Optimization variables

The resource allocation optimization problem is supposed to cost-efficiently select the values of the optimization variables $x_{i,s,t}$, $y_{i,s,k,l}$, and $f_{i,s,m}$ to determine the functional split, midhaul lightpath, and reliability measure for each network slice. The binary variables $x_{i,s,t}$ equals 1 if the functional split t is assigned to slice s in ith DU. $x_{i,s,t}, t = 0, 1, 2$ corresponds to the three functional splits described in Section II, while $x_{i,s,3}$ relates to the generalized MEC functional split. The binary variable $y_{i,s,k,l}$ takes 1 if the aggregated traffic of slice s in DU i is routed over lth wavelength of the kth available path from DU i to CU. $P_{i,k}$ denotes kth available path from DU i to the CU. The reliability measure for the sth slice of ith DU are determined by the binary variable $f_{i,s,m}$, where $m = 0$ means neither FEC nor packet duplication is used, $m = 1, 2, 3$ correspond to using FEC level m without packet duplication, $m = 4$ associates with pure packet duplication without FEC, and finally $m = 5, 6, 7$ imply that FEC level $m-4$ with packet duplication is used. The auxiliary variable $p_{i,s} = \sum_{m=4}^{7} f_{i,s,m}$ shows whether the slice s of DU i uses packet duplication or not. Further, the auxiliary variable $w_{i,s,k}$ means whether $P_{i,k}$ is chosen for slice s of DU i or not. Moreover, the auxiliary variables $z_{i,s,t,m}$ and $g_{i,sk}$ are used to linearize constraints.

B. Optimization constraints

The constraints of the optimization problem can be classified into three categories of physical, capacity, and QoS constraints. Each category is described as follows.

1) Physical constraints: Each slice s of DU i should choose one functional split enforced by

$$\sum_{t=0}^{3} x_{i,s,t} = 1 \quad \forall i, s. \tag{23a}$$

Only slices eligible for using MEC can select the generalized MEC functional split, as given by

$$x_{i,s,3} \leqslant M_{i,s} \quad \forall i, s \tag{23b}$$

TABLE II: List of indices, parameters, and variables along with their corresponding names. $\mathbb{N}$, $\mathbb{W}$, and $\mathbb{R}$ are sets of natural, whole, and real numbers, respectively. $\mathbb{X}^c$ means all numbers in set $\mathbb{X}$ that satisfies condition c.

Type	Notation	Name
Indices	$i \in \mathbb{N}^{\leqslant N}$	Node index
	$j \in \mathbb{N}^{\leqslant F}$	Link index
	$s \in \mathbb{N}^{\leqslant S_i}$	Slice index
	$t \in \mathbb{W}^{\leqslant 3}$	Split index
	$l \in \mathbb{N}^{\leqslant L}$	Wavelength index
	$k \in \mathbb{N}^{\leqslant P_i}$	Path index
	$m \in \mathbb{W}^{\leqslant 7}$	Reliability index
Parameters	$P_i \in \mathbb{N}$	Number of paths between DU i and CU
	$N \in \mathbb{N}$	Number of DU nodes
	$F \in \mathbb{N}$	Number of fiber links
	$L \in \mathbb{N}$	Number of wavelengths
	$S \in \mathbb{N}$	Maximum number of slices in a DU node
	$S_i \in \mathbb{N}^{\leqslant S}$	Number of slices in DU i
	$C_i \in \mathbb{R}^{\geqslant 0}$	Available processing capacity in DUs
	$C_0 \in \mathbb{R}^{\geqslant 0}$	Available processing capacity in CU
	$E \in \mathbb{R}^{\geqslant 0}$	Overall network cost
	$E^B \in \mathbb{R}^{\geqslant 0}$	Total BPF processing cost
	$E^F \in \mathbb{R}^{\geqslant 0}$	Total FEC processing cost
	$E^N \in \mathbb{R}^{\geqslant 0}$	Total connectivity cost
	$A \in \mathbb{R}^{\geqslant 0}$	Equivalent transmission cost
	$M_{i,s} \in \mathbb{R}^{\geqslant 0}$	MEC possibility
	$U_{i,k,j} \in \{0,1\}$	Link membership indicator
	$D_{i,k,t,m} \in \mathbb{R}^{\geqslant 0}$	Achievable delay
	$R_{i,k,m} \in (0,1]$	Achievable PER
	$T_{i,s} \in \mathbb{R}^{\geqslant 0}$	Tolerable delay
	$H_{i,s} \in (0,1]$	Tolerable PER
	$O_{i,s,t} \in \mathbb{R}^{\geqslant 0}$	Required BPF processing capacity at DU
	$Q_{i,s,t} \in \mathbb{R}^{\geqslant 0}$	Required BPF processing capacity at CU
	$V_{i,s,m} \in \mathbb{R}^{\geqslant 0}$	Required FEC processing capacity
Variables	$x_{i,s,t} \in \{0,1\}$	Functional split selector
	$y_{i,s,k,l} \in \{0,1\}$	Wavelength-path selector
	$f_{i,s,m} \in \{0,1\}$	Reliability measure selector
	$p_{i,s} \in \{0,1\}$	(auxiliary) Packet duplication selector
	$w_{i,s,k} \in \{0,1\}$	(auxiliary) Path selector
	$z_{i,s,t,m} \in \{0,1\}$	(auxiliary) $f_{i,s,m}x_{i,s,t}$
	$g_{i,s,k} \in \{0,1,2\}$	(auxiliary) $w_{i,s,k}(1+p_{i,s}-x_{i,s,3})$

, where $M_{i,s}$ is a constant parameter taking 0 if MEC processing is not allowed for slice s of DU i, and 1 otherwise. The constraint

$$\sum_{k=0}^{P_i}\sum_{l=0}^{L} y_{i,s,k,l} = 1 + p_{i,s} - x_{i,s,3} \quad \forall i,s \tag{23c}$$

sets the number of lightpaths connecting slice s of DU i to the CU. Normally, a single lightpath is required for each slice, unless the slice uses MEC and is processed in the DU, where no lightpath is required, or the slice uses packet duplication in which two wavelengths over a same path are required. To force a single path is selected for the two wavelengths involved in packet duplication,

$$\sum_{k=0}^{P_i} w_{i,s,k} = 1 \quad \forall i,s \tag{23d}$$

$$, \quad \sum_{l=0}^{L} y_{i,s,k,l} = g_{i,s,k} \quad \forall i,s,k \tag{23e}$$

, where the first constraint imposes the selection of one path for each slice, and the second constraint employs the auxiliary integer variable $g_{i,s,k}$ to assure that the proper number of wavelengths are reserved over the path. $g_{i,s,k} = w_{i,s,k}(1 + p_{i,s} - x_{i,s,3})$ is provided by the linear constraints

$$\begin{aligned} & 2w_{i,s,k} + (1 + p_{i,s} - x_{i,s,3}) - 2 \leqslant g_{i,s,k} \quad \forall i,s,k \\ , \quad & g_{i,s,k} \leqslant 2w_{i,s,k} \quad \forall i,s,k \\ , \quad & g_{i,s,k} \leqslant 1 + p_{i,s} - x_{i,s,3} \quad \forall i,s,k. \end{aligned} \tag{23f}$$

No FEC and packet duplication is required when the slice is locally processed at MEC server, as constrained by

$$f_{i,s,0} = x_{i,s,3} \quad \forall i,s. \tag{23g}$$

Selecting a single reliability measure is assured by

$$\sum_{m=0}^{7} f_{i,s,m} = 1 \quad \forall i,s. \tag{23h}$$

2) Capacity constraints: The processing tasks performed in each DU should be kept below its available processing capacity defined as

$$\sum_{s=0}^{S_i}\Big[\sum_{t=0}^{3} x_{i,s,t}O_{i,s,t} + \sum_{m=0}^{7} f_{i,s,m}V_{i,s,m}\Big] \leqslant C_i \quad \forall i \tag{24a}$$

, where the first and second terms correspond to the resources required for BPF and FEC processing, respectively, and C_i denotes the available processing capacity in DU i. The capacity constraint for the CU is defined as

$$\sum_{i=0}^{N}\sum_{s=0}^{S_i}\Big[\sum_{t=0}^{3} x_{i,s,t}Q_{i,s,t} + \sum_{m=0}^{7} f_{i,s,m}V_{i,s,m}\Big] \leqslant C_0 \tag{24b}$$

, where the first and second terms are related to the total processing resources needed for BPF and FEC processing in the CU, respectively, and C_0 is the CU processing capacity. A single wavelength should be continuously used for each lightpath of the midhaul network given by the constraint

$$\sum_{i=0}^{N}\sum_{s=0}^{S_i}\sum_{k=0}^{P_i} y_{i,s,k,l}U_{i,k,j} \leqslant 1 \quad \forall j,l \tag{24c}$$

, where $U_{i,j,k}$ is a binary parameter which is 1 if link j is used in path k from DU i to the CU.

3) QoS constraints: The latency and reliability requirement of the slices should be guaranteed. Let $D_{i,k,t,m}$ be the achievable delay if a slice of DU i with functional split t and reliability measure m is served over path $P_{i,k}$ and define $T_{i,s}$ as the maximum tolerable delay of slice s. Delay requirements are maintained by

$$y_{i,s,k,l} \leqslant 1 - z_{i,s,t,m} \quad \forall i,s,k,l,m : T_{i,s} \leqslant D_{i,k,t,m} \tag{25a}$$

, where $z_{i,s,t,m}$ is an auxiliary binary variable showing that functional split $x_{i,s,t}$ and reliability measure $f_{i,s,m}$ are chosen. Indeed, $z_{i,s,t,m} = x_{i,s,t}f_{i,s,m}$, which can be equivalently expressed by the linear constraints

$$\begin{aligned} & x_{i,s,t} + f_{i,s,m} - 1 \leqslant z_{i,s,t,m} \quad \forall i,s,m,t \\ , \quad & z_{i,s,t,m} \leqslant x_{i,s,t} \quad \forall i,s,m,t \\ , \quad & z_{i,s,t,m} \leqslant f_{i,s,m} \quad \forall i,s,m,t. \end{aligned} \tag{25b}$$

Let $R_{i,k,m}$ be the achievable PER corresponding to the selection of reliability measure $f_{i,s,m}$ over path $P_{i,k}$. Further, define $H_{i,s}$ as the tolerable PER of slice s in DU i. The reliability requirement is guaranteed by the constraint

$$y_{i,s,k,l} \leqslant \sum_{\{m=0,\cdots,7 | R_{i,k,m} < H_{i,s}\}} f_{i,s,m} \quad \forall i,s,k,l \tag{25c}$$

, which avoids the selection of the lightpaths incapable of providing the tolerable PER $H_{i,s}$.

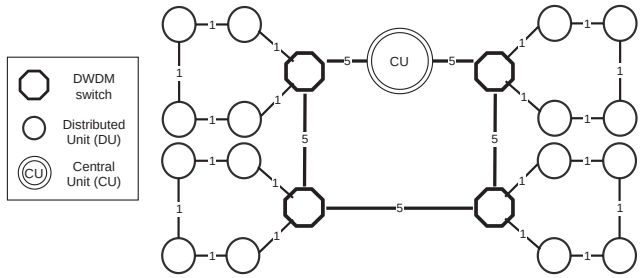

Fig. 1: Benchmark network topology with $N = 20$ nodes and $F = 25$ links. The numbers over the links represent their length in km.

C. Objective Function

The resource allocation targets the minimization of the overall network cost E including total costs of BPF processing E^B, FEC processing E^F, and connectivity E_N as

$$\min_{\substack{x_{i,s,t},\, f_{i,s,m},\, y_{i,s,k,l} \\ p_{i,s},\, w_{i,s,k},\, g_{i,s,k},\, z_{i,s,t,m}}} \quad E = E^B + E^F + E^N. \tag{26a}$$

E^B and E^F equal

$$E^B = \sum_{i=0}^{n} \sum_{s=0}^{S_i} \sum_{t=0}^{3} x_{i,s,t}\big[O_{i,s,t} + Q_{i,s,t}\big] \tag{26b}$$

$$, \quad E^F = \sum_{i=0}^{n} \sum_{s=0}^{S_i} \sum_{m=0}^{7} f_{i,s,m} V_{i,s,m}. \tag{26c}$$

In (26b) and (26c), the first term is the sum of processing costs in DUs and the second term corresponds to processing cost in CU, where the cost is measured in terms of the required processing capacity. The connectivity cost C^N is

$$E^N = A \sum_{i=0}^{n} \sum_{s=0}^{S_i} \big[1 + p_{i,s} - x_{i,s,3}\big] \tag{26d}$$

, where the number of used lightpaths is multiplied to the transmission cost A. We assume that the transmission on a lightpath of the midhaul network costs as much as the cost of occupying A processing resources in the CU. Clearly, the FECs are processed as VNFs with the cost described by (26c). So, FEC processing cost is excluded from the connectivity cost given in (26d).

IV. Numerical Results

The ILP formulation has $O(NSL \max_i\{P_i\})$ variables and $O(NSL \max_i\{P_i\})$ constraints, where $O(.)$ denotes big-O notation. Fortunately, the linear structure of the formulation makes solving the problem affordable for many practical scenarios. In this section, the performance of the proposed resource allocation scheme is evaluated via simulation for several sample scenarios. The optimization problem is implemented in YALMIP [16], numerically solved by CPLEX [17], and analytically investigated using MATLAB. We use the OTN/DWDM midhaul network topology of Fig. 1 with $L = 20$ devoted 50-GHz wavelengths for simulations. Each wavelength provides a transmission rate of 10 Gbps, enough to carry the data stream of each slice.

The number of slices S_i in DU i is chosen uniformly from the integers $0, 1, \cdots, S$, where S specifies the maximum number of slices in each DU. The aggregated traffic $\lambda_{i,s}$ of slice s in DU i is derived from a trimmed normal distribution with mean Λ Gbps and variance 1. Moreover, the maximum acceptable reliability and latency as well as MEC possibility of slice s in DU i are randomly chosen from the three options below.

TABLE III: Simulation parameters and their values.

Parameter	Value	Parameter	Value
Fiber attenuation	0.22 dB/km	Fiber delay	5 μs/km
Two-degree switching loss	3 dB	Switching delay	5 μs
Multi-degree switching loss	11 dB	Spectral efficiency	1 bit/Hz/s
Noise temperature	300 K	Load resistor	50 Ω
Spontaneous emission factor	1.58	Dark current	5 nA
Optical bandwidth	50 GHz	Electrical bandwidth	10 GHz
Quantum efficiency	0.75	Transmit power	0 dBm
Average packet length	1000 bit	Working bandwidth	155 nm

1) $H_{i,s} \leqslant 10^{-4}$, $T_{i,s} \leqslant 5$ ms, $M_{i,s} = 0$ with a probability of 50%.
2) $H_{i,s} \leqslant 10^{-5}$, $T_{i,s} \leqslant 500$ μs, $M_{i,s} = 0$ with a probability of 25%.
3) $H_{i,s} \leqslant 10^{-5}$, $T_{i,s} \leqslant 500$ μs, $M_{i,s} = 1$ with a probability of 25%.

The first option characterizes a delay-tolerant normal slice while the two other options describe a delay-sensitive reliable slice with or without MEC possibility.

In a typical air interface configuration, the reference core (RC) characterized in [18] can afford serving BPFs of the full radio protocol stack within a one-way latency of $\tau^B_{i,s} = 1$ ms. Scaling the results of [18], $17.2 = o_{i,s,t} + q_{i,s,t}$ RCs are roughly required to process the full protocol stack for a 1 Gbps delay-tolerable normal slice, where $o_{i,s,0} = 0$, $o_{i,s,1} = 14.1$, $o_{i,s,2} = 15.6$, and $o_{i,s,3} = 17.2$ RCs, respectively. Depending on the selected functional split t, $o_{i,s,t}$ of the required RCs are located in DU i while the remaining $q_{i,s,t}$ RCs sit in the CU [18]. Generally, processing BPFs of a delay-tolerable normal slice with data rate $\lambda_{i,s}$ costs $O_{i,s,t} = \alpha_i \lambda_{i,s} o_{i,s,t}$ and $Q_{i,s,t} = \lambda_{i,s} q_{i,s,t}$ on DU i and CU, respectively, where $\alpha_i = 2$ stands for relative cost of occupying one RC in DU i with respect to that of in CU. For a delay-sensitive reliable slice, the involved BPFs can be roughly resolved within a one-way latency of $\tau^B_{i,s} = 200$ μs at the cost of occupying $5 \times 17.2 = o_{i,s,t} + q_{i,s,t}$ RCs, where an ideal parallel processing gain of 5 is applied. The same processing capacity of $C_i = C$ RCs is deployed in each DU while $C_0 = 1000$ RCs is available in the CU.

A native software implementation of ITU G.709.2 standard on the considered RC is used to approximate the required capacity $V_{i,s,m} = (1+\alpha_i)(62.5+14m)$ RCs, processing delay $\tau^F_m = 26m$ μs, and PER reduction capability $\eta_m = 10^{-m}$ (for pre-FEC errors $[10^{-3}, 10^{-5}]$) of the staircase FECs for different reliability measures $m = 1, 2, 3$, each having its own decoding window length. The coefficient $1+\alpha_i$ in $V_{i,s,m}$ stands for the FEC processing cost at the source and destination of a lightpath at DU i and CU. Clearly, when no FEC is used $V_{i,s,0} = 0$, $\tau^F_0 = 0$, and $\eta_0 = 1$. For the reliability measures $m = 4, 5, 6, 7$, packet duplication along with a same FEC as reliability measure $m - 4$ is used, so $V_{i,s,m} = V_{i,s,m-4}$, $\tau^F_m = \tau^F_{m-4}$, and $\eta_m = \eta_{m-4}$ for $m = 4, 5, 6, 7$. The paths $P_{i,k}$ are obtained by computing all available paths from node i to CU and then, $U_{i,k,j}$ are determined. The achievable latency $D_{i,k,t,m}$ and reliability $R_{i,k,m}$ over path $P_{i,k}$ are determined in an offline pre-computation stage. Achievable delay $D_{i,k,t,m} = \tau^P_{i,k} + \tau^B_{i,s} + \tau^F_m$, where the path delay $\tau^P_{i,k}$ includes the propagation and switching delays over path $P_{i,k}$. We assume that there is only a pre-amplification stage at the binary direct detection (DD) receiver of each lightpath, with-

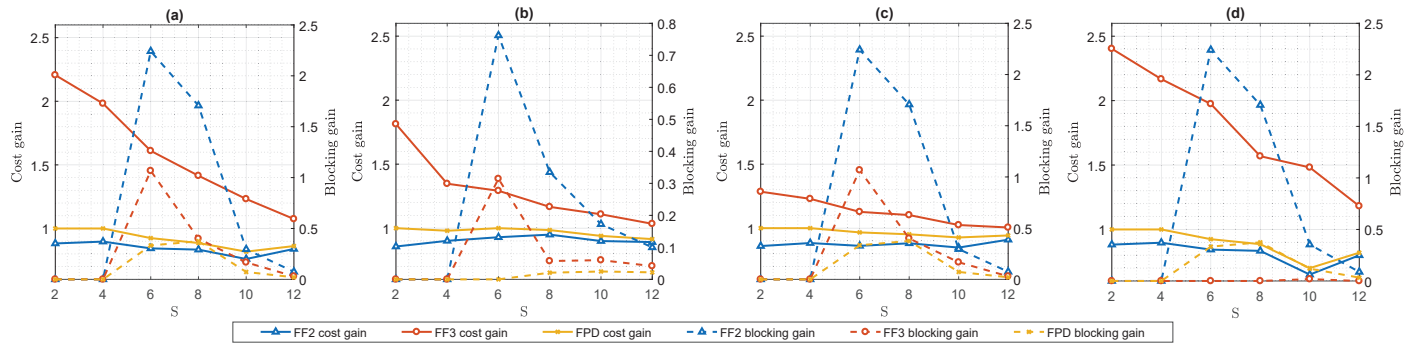

Fig. 2: Cost gain and blocking gain with various benchmark schemes versus maximum number of slices S in four distinguished scenarios. (a) $C = 200$, $A = 10$, $\Lambda = 0.2$ (b) $C = 200$, $A = 10$, $\Lambda = 1$ (c) $C = 200$, $A = 50$, $\Lambda = 0.2$ (d) $C = 600$, $A = 10$, $\Lambda = 0.2$.

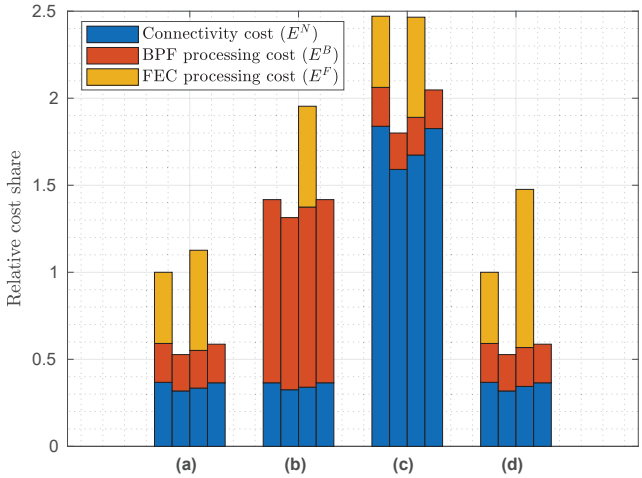

Fig. 3: Relative cost share of the considered schemes in four distinguished scenarios with $S = 6$. (a) $C = 200$, $A = 10$, $\Lambda = 0.2$ (b) $C = 200$, $A = 10$, $\Lambda = 1$ (c) $C = 200$, $A = 50$, $\Lambda = 0.2$ (d) $C = 600$, $A = 10$, $\Lambda = 0.2$. In each group, the leftmost bar represents the DRM scheme, the next two bars represent the FF2 and FF3 schemes, respectively, and the rightmost bar represents FPD scheme. All bars are normalized to the value of the leftmost bar with a relative overall cost of 1.

out any intermediate optical-electrical conversion or optical amplification, and use equations (8.1.14) and (8.3.1) of [19] to compute the received pre-FEC PER $\epsilon_{i,k}$. The achievable PER is obtained by $R_{i,k,m} = \epsilon_{i,k}\eta_m, m = 0, 1, 2, 3$ and $R_{i,k,m} = \epsilon^2_{i,s,k}\eta_m, m = 4, 5, 6, 7$, where we have assumed that the packet duplication reduces the error rate by square of the received pre-FEC PERs. Tab. III summarizes the constant parameters used in calculation of $D_{i,k,t,m}$ and $R_{i,k,m}$.

The performance of the proposed scheme is compared with several benchmark schemes including two fixed FEC schemes, called FF2 and FF3, and one fixed packet duplication scheme abbreviated as FPD. In the schemes FF2 and FF3, the selection of the first and second FEC configuration is respectively forced while packet duplication without any FEC is compelled in the FPD scheme. Since any reliability measure can be arbitrarily selected from the available options, the proposed scheme is referred to as dynamic reliability measure (DRM). We report the performance metrics of cost gain and blocking gain for various values of S, C, and A in different working scenarios. The cost gain is defined as the ratio of the overall network cost of a benchmark scheme to that of the DRM scheme. The blocking gain is defined as the ratio of the difference between minimum number of blocked slices required for a convergent resource allocation optimization in a benchmark and the DRM scheme to that of in the DRM scheme. Cost gain measures how the flexible nature of the DRM reduces the overall cost while blocking gain quantifies the capability of the DRM scheme in efficient utilization of network capacity . Each performance metric value is obtained by taking an average over 10 simulation runs, each one corresponding to an independent realization of the considered random parameters.

Fig. 2 reports the cost and blocking gains for different benchmark schemes versus the maximum number of slices S in various working scenarios. Cost of benchmark scenarios can be as high as two times the cost of the proposed DRM. The involved synergy between optical and radio segments in the proposed scheme allows to use a cost-effective reliability measure to guarantee the reliability commitments while assuring the latency requirements by a proper combined selection of the functional split and reliability measure to reduce processing delay. As the reference schemes are more constrained than the DRM scheme, they experience higher request blocking and higher overall cost. However, with higher blocking gains, benchmark schemes do not pay for the resources required to serve blocked slices while the DRM does at the expense of the lower cost gain, as can be interpreted from Fig. 2(a). As can be seen in Fig. 2(b), increasing the traffic load Λ reduces blocking gain since more slices are blocked in the DRM scheme due to higher occupancy of the resources in DUs. When $S = 6$, a sudden peak in the blocking gain is observed since the DRM scheme still affords serving the slices while the benchmark schemes have to block some requests. When the transmission cost A increases in Fig. 2(c), the cost gain compared to the FF3 scheme declines since the cost-efficiency of packet duplication, which mostly uses connectivity resources, reduces. If more processing capacity is provided in DUs, FF3 scheme will have enough resources for the FEC processing to reduce the number of blocked slices with stringent reliability requirements. As shown in Fig. 2(d), consuming more processing resources in FF3 scheme improves the cost gain of the DRM scheme which utilizes the available resource more cost-effectively. Despite increasing the processing capacity of DUs in Fig. 2(d), the scheme FF2 still suffers from high blocking rates because its rigid FEC configuration provides a fixed PER reduction capability of 0.01, which is not enough to afford the stringent reliability requirements of the slices without MEC possibility.

Fig. 3 reports the relative cost share of the four considered scenarios of Fig. 2 for $S = 6$. It can be observed that increasing Λ and A, share of BPF processing cost E^B and connectivity cost E^N grow up, as shown in Figs. 3(b) and 3(c), respectively. The equal size of connectivity cost bars for DRM and FPD schemes in all four scenarios shows that the DRM favors

packet duplication over FEC since the FEC processing cost is higher than connectivity cost even with $A = 50$. As stated before, providing more processing capacity in DUs allows more FEC processing in FF3 scheme, which increases the cost share of FEC processing in Fig. 3(d).

V. Conclusion

In this paper, we propose a combined resource allocation and functional split selection scheme to show how a constructive synergy between optical and radio segments of a RAN built on OTN/DWDM network can bring cost saving. Particularly, an optimization problem for joint allocation of connectivity and processing resources is formulated to optimally select function splits and transmission configurations such that the overall cost is minimized constrained to the satisfaction of physical and capacity limitations as well as reliability and latency requirements. The service reliability requirement can be guaranteed by using FEC processing; however, when the processing capacity is not enough, the reliability may be obtained by packet duplication at the cost of consuming more connectivity resources. Although packet duplication may incur higher cost, it helps to reduce request blocking and utilize the network more efficiently. The flexible selection of the FEC level and packet duplication in the proposed DRM scheme provides remarkable cost gains with considerably lower request blocking compared to the benchmark schemes.

References

[1] W. Ejaz *et al.*, "A comprehensive survey on resource allocation for cran in 5G and beyond networks," *Journal of Network and Computer Applications*, vol. 160, p. 102638, 2020.

[2] Y. Xiao *et al.*, "Can fine-grained functional split benefit to the converged optical-wireless access networks in 5G and beyond?" *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1774–1787, Sep. 2020.

[3] Y. Li *et al.*, "Flexible RAN: Combining dynamic baseband split selection and reconfigurable optical transport to optimize ran performance," *IEEE Network*, vol. 34, no. 4, pp. 180–187, 2020.

[4] P. Marsch, Ö. Bulakci, O. Queseth, and M. Boldi, *5G system design: architectural and functional considerations and long term research*. John Wiley & Sons, 2018.

[5] F. W. Murti *et al.*, "On the optimization of multi-cloud virtualized radio access networks," in *International Conference on Communications (ICC)*. IEEE, 2020.

[6] B. Ojaghi *et al.*, "Sliced-ran: Joint slicing and functional split in future 5G radio access networks," in *International Conference on Communications (ICC)*. IEEE, 2019.

[7] A. Garcia-Saavedra *et al.*, "Joint optimization of edge computing architectures and radio access networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2433–2443, 2018.

[8] S. Perrin, "5G transport networks: Heavy reading operator survey & analysis," *Heavy Reading*, 2018.

[9] M. Klinkowski, "Latency-aware DU/CU placement in convergent packet-based 5G fronthaul transport networks," *Applied Sciences*, vol. 10, no. 21, p. 7429, 2020.

[10] B. M. Khorsandi *et al.*, "Dedicated path protection for reliable network slice embedding based on functional splitting," in *International Conference on Transparent Optical Networks (ICTON)*. IEEE, 2019.

[11] Y. Li *et al.*, "End-to-end urllc slicing based on packet duplication in 5G optical transport networks," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 12, no. 7, pp. 192–199, 2020.

[12] G. Mountaser *et al.*, "Reliable and low-latency fronthaul for tactile internet applications," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2455–2463, 2018.

[13] "OTU4 long-reach interface," *Recommendation G.709.2, International Telecommunication Union (ITU)*, Oct. 2018.

[14] B. P. Smith *et al.*, "Staircase codes: Fec for 100 Gb/s OTN," *Journal of Lightwave Technology*, vol. 30, pp. 110–117, 2012.

[15] B. S. G. Pillai *et al.*, "End-to-end energy modeling and analysis of long-haul coherent transmission systems," *Journal of Lightwave Technology*, vol. 32, no. 18, pp. 3093–3111, 2014.

[16] J. Lofberg, "Yalmip: A toolbox for modeling and optimization in matlab," in *2004 IEEE international conference on robotics and automation (IEEE Cat. No. 04CH37508)*. IEEE, 2004, pp. 284–289.

[17] C. U. Manual, "Ibm ilog cplex optimization studio," *Version*, vol. 12, pp. 1987–2018, 1987.

[18] N. Nikaein, "Processing radio access network functions in the cloud: Critical issues and modeling," in *International Workshop on Mobile Cloud Computing and Services*, 2015, pp. 36–43.

[19] R. Hui, *Introduction to fiber-optic communications*. Academic Press, 2019.