This is a postprint version of the following published document :

Marotta, A., Cassioli, D., Kondepu, K., Antonelli, C. y Valcarenghi, L. (2019). Exploiting flexible functional split in converged software defined access networks. *Journal of Optical Communications and Networking,* 11(11), pp. 536-546.

DOI: https://doi.org/10.1364/JOCN.11.000536

Exploiting Flexible Functional Split in Converged Software Defined Access Networks

Andrea Marotta, *Student Member, IEEE,* Dajana Cassioli, *Senior Member, IEEE,*
Koteswararao Kondepu, *Senior Member, IEEE,* Cristian Antonelli, *Senior Member, IEEE,*
Luca Valcarenghi, *Senior Member, IEEE,*

***Abstract*—5G targets to offer a huge network capacity to support the expected unprecedented traffic growth mainly due to mobile and machine-type services. Thus 5G access network has to comply with very challenging architectural requirements. Mobile networks scalability is achieved by playing appropriately with the centralization of network functions and by applying the functional split introducing the fronthaul. Although more advantageous in terms of network management and performance optimization, low layer functional split options require larger bandwidth and lower latency to be guaranteed by the fronthaul in the access network, while preserving other concurrent FTTx services. Thus, advanced mechanisms for the efficient management of available resources in the access network are required to jointly control both radio and optical domains. Softwarized mobile and optical segments facilitate the introduction of dedicated protocols to enable the inter-working of the two control planes. This paper proposes a new cooperation scheme to manage the adaptive flexible functional split in 5G networks conditioned to the resource availability in the optical access network. Techniques for the accurate estimation of available bandwidth and the associated real-time selection of the best suitable functional split option are investigated. Results show that the proposed software defined converged approach to wavelength and bandwidth management guarantees the optimal allocation of optical resources. Triple exponential smoothing forecasting technique enables efficient coexistence of mobile fronthaul and fixed connectivity traffic in the network, reducing traffic impairments with respect to other well-known forecasting techniques, while keeping the same level of centralization.**

***Index Terms*—Flexible Functional Split; Fronthaul; 5G; TWDM-PON; Software Defined Networks.**

I. Introduction

The 5G concept as a *Network of Networks* [1] poses several challenges in terms of effective interoperability of different layers and domains of the underlying tangled architecture, and network management approaches become crucial. An unprecedented degree of flexibility is required to enhance the responsiveness of the network to the instantaneous traffic load with heterogeneous requirements for Quality of Service/Experience (QoS/QoE).

The highest level of flexibility is nowadays granted by *network softwarization*, where the network architecture can be adapted to the instantaneous requirements, and the delivery of "network functionality via software running on industry-standard commercial off-the-shelf (COTS) hardware" and the programmability of network entities are enabled [2], [3].

Such flexibility provides the opportunity to split the network functionalities over several different planes. The simplest one is the "vertical functional split", offered by the concept of Software Defined Networks (SDNs). It allows the separation of control plane and user plane providing a *logically centralized control*. Besides, the centralization can be applied at the level of virtualized network functions and it is referred to as "horizontal functional split," which enables to vary the centralization degree. The horizontal functional split, hereafter is referred to as *Functional Split* (FS).

Recently, different possible FS options have been proposed [4]–[6], where low layer FSs provide the highest centralization degree. Each of these options determine specific latency and capacity requirements on the fronthaul, for which numerous transport interfaces are currently under investigation and standardization. The centralization brings several advantages in mobile networks, especially in terms of coordination for dynamic resource allocation and latency minimization [7]. On the other hand, low layers FSs impose very strict latency constraints that do not allow data transport over a long distance [8].

The allocation of virtualized mobile network functions can be flexibly operated in the central unit (CU), in the distributed units (DUs), or in the Radio Units (RUs) (i.e. closer to the antenna sites [9]) in a cloud radio access network (C-RAN). Different mappings of CU/DU/RU functions on split architectures composed of either two or three elements are described in [10]. In [11], [12], it has been presented that except for analog Radio Frequency and Digital-Analog/Analog-Digital conversion all the other functions (i.e., from split Option 8 above) can be implemented in software through virtual machines and containers. As a proof, the following work showed a successful implementation of 5G functions in containers [13]. Thus, when split option is changed, only software updates/activation are needed.

A promising approach to relax the excessive fronthaul requirements is the *Flexible Functional Split* (FFS), which is widely studied in the recent literature and standardization efforts are currently in progress [14]. By FFS the network functionalities can be distributed dynamically and flexibly between network units [15], [16], as experimentally demonstrated in [7] and [17]. According to 3GPP [18], FFS should provide the ability to adapt to transport network performance

A. Marotta, D. Cassioli, and C. Antonelli are with University of L'Aquila, L'Aquila, Italy (e-mail: andrea.marotta@univaq.it; dajana.cassioli@univaq.it; cristian.antonelli@univaq.it).

K. Kondepu and L. Valcarenghi are with Scuola Superiore Sant'Anna, Pisa, Italy (e-mail: k.kondepu@santannapisa.it; l.valcarenghi@santannapisa.it.

level. Thus, it allows to reduce costs related to optical access network deployment that would be considerably higher in the case of fixed functional split requiring fixed network performance level.

The dynamic allocation of network functionalities on network units requires the deployment of hardware resources to host the functionalities on both sides, which could appear as a waste of resources, with consequent increase of CApital EXpenditures (CAPEX) for operators. However, it is expected that these resources, when not utilized by the fronthaul, could be profitably used for different purposes or services like e.g. caching, mobile cloud computing, clone cloud, etc. [19]. The decision of the instantaneous best suitable FFS option to be dynamically adopted in the mobile network is made based on merit functions calculated on specific key performance indicators (KPIs) and parameters, like, e.g., the inter-cell interference and fronthaul bandwidth utilization [20].

Moreover, Passive Optical Networks (PONs) have been proposed as possible fronthaul infrastructures thanks to the low OPerational EXpenditures (OPEX) and high availability in urban scenarios [21]. Often, the fronthaul shares the resources available in the optical access network with fixed connectivity traffic. Thus, appropriate mechanisms for multiple access to optical resources are required to meet bandwidth and latency constraints of all services to be supported [22]. We distinguish between these two types of traffic: the fronthaul traffic and the traffic generated by fixed connectivity services such as Fiber To The X (FTTx) residential and business connectivity. An ultra low latency medium access control mechanism capable of supporting CPRI-based mobile fronthaul latency requirements is demonstrated in [23] utilizing a commercial Time-Division Multiplexed Passive Optical Network (TDM-PON) platform. The proposed scheme is validated in FTTx coexistence conditions.

In addition, significant advantages are provided by converged management schemes over different domains. Actually, the optical access domain can make use of Software Defined Access (SDA) with special focus on the integration between Access and Aggregation networks [24]. Core and Radio Access Networks relying on Software Defined Mobile Network (SDMN) approach can apply an advanced joint management of resources, spectrum and mobility, while boosting the cooperation among heterogeneous networks [25].

Converged Dynamic Bandwidth Allocation (DBA) of mobile fronthaul over TDM-PON can be realized by allocating the time slots in the TDM-PON in cooperation with the mobile scheduler based on the estimated data arrival time [26], [27]. However, this approach requires an ad-hoc dedicated interface between the CU and the OLT for forwarding the scheduling information.

In [28], a novel SDA and SDMN integrated framework for a PON-based fronthauling solution has been investigated, where the devised Software Defined Wavelength Bandwidth Allocation (SD-WBA) scheme brings a significant gain in terms of both performance and cost. However, the investigation was limited to only a fixed functional split option.

In this paper, we define a new cooperation scheme for the inter-working of the SDA and SDMN controllers where the fast re-configuration and distribution of network functions between CU/DU and RU/DUs in the mobile network is driven by the SD-WBA management scheme. The proposed approach "modulates" the fronthaul traffic required by the mobile segment by selecting the FS option based on the resources available in the Time and Wavelength Division Multiplexing PON (TWDM-PON) once the bandwidth has been allocated to fixed connectivity services.

Moreover, the paper also introduces a bandwidth forecasting technique for the fixed connectivity traffic to change a FS based on the accurate estimation of resources made available to the fronthaul between two consecutive FS reconfigurations. In our previous study [29], a FFS without forecasting has been studied, here, we consider and compare different techniques for bandwidth forecasting to achieve efficient coexistence of fronthaul and fixed connectivity traffic such as FTTx and business connectivity.

II. Reference Architecture

The considered reference architecture, shown in Fig. 1, consists of a urban scenario where several deployed network units are connected through a TWDM-PON.

We distinguish between DU/RU units equipped with antenna which are able to host RLC, MAC, PHY and RF functions, and CU/DU units which host higher layers functionalities. Both CU/DU and DU/RU units run over commodity hardware on which virtual functions are dynamically deployed in order to implement FS options listed in Table I. As shown in Fig. 1, different FS options correspond to different configurations of mobile network units. Furthermore, based on the implemented functional split different logical interfaces (i.e. CPRI, Fx, F1) are adopted for signaling exchange and data transmission between mobile network units.

The different configurations of CU/DU and DU/RU are assumed to be virtual functions pre-stored at mobile network units. Thus, no transfer of software blocks is required for the functional split reconfiguration. Moreover, the virtualization of the RAN components is experimentally demonstrated with different types of hypervisor technologies (e.g., VirtualBox, Docker) in [13].

Each DU/RU is attached to an Optical Network Unit (ONU), whereas the CU/DU is connected to the Optical Line Termina-

Table I: Split Options requirements [10], [11]

Split option	UL [Mbps]	DL [Mbps]	One-Way Max delay	DU/RU Configuration	CU/DU Configuration	Supported Interface
2	74	166	1.5˜10 ms	RU	Colocated CU and DU	F1
6	119	211	100-500 μs	RU	Colocated CU and DU	Fx
7	674.4	479	100-500 μs	RU	Colocated CU and DU	Fx
8	1966	1966	$\leq$100 μs	Colocated DU and RU	CU	CPRI

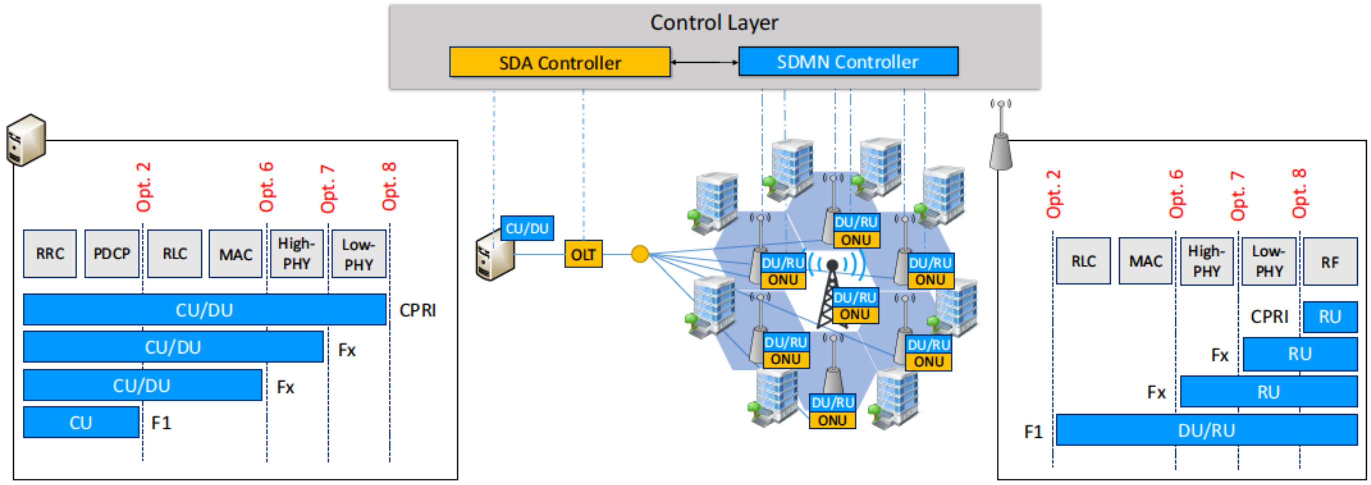


Figure 1: Reference architecture.

tion (OLT) at the central office site. The PON also constitutes the infrastructure to provide fixed connectivity such as residential and business connectivity. The optical infrastructure is managed by an SDA controller which interacts with the agents placed at the OLT and the ONUs. The SDA controller is responsible of the Dynamic Wavelength and Bandwidth Allocation (DWBA), wavelength activation/deactivation and integrated QoS management with the metro network. Moreover, it exposes relevant parameters of the optical access network to other SDN Controllers through the so called *west-bound* interface that allows convergence among different control planes [30].

The SDMN controller is in charge of managing the mobile network. It is responsible of Radio Resource Management (RRM), cells activation/deactivation, cooperation among mobile network units and mobile network function placement. Similarly to the SDA controller, the SDMN controller enables integrated mobile-optical control mechanisms by means of the west-bound interface.

The exchange of information between optical and mobile controllers is exploited in the following two steps: (i) the dynamic adaptation of the FS option in the fronthaul based on the available bandwidth after accommodating the fixed connectivity traffic; (ii) the application of the SD-WBA in the PON to support fronthauling. As already mentioned, the different FS options imply different requirements for the fronthaul segment in terms of peak data-rate and delay budget as reported in Table I. The values in Table I are calculated according to [11], and by considering a 20MHz MIMO 2x2 system with 64 QAM modulation format in downlink and 16 QAM modulation format in uplink. The overhead values adopted in the following are obtained from [31]. For example, option 2 bandwidth requirement for uplink/donwlink $B_{2,U/D}$ is calculated as in Eq. (1). We use the notation $B_{2,U/D}$ to denote two variables $B_{2,U}$ and $B_{2,D}$ representing respectively uplink and downlink bandwidth and this notation is utilized throughout the section.

$$B_{2,U/D} = R_{U/D} + O_{PR,U/D} \qquad (1)$$

where:

- $R_{U/D}$ is the LTE uplink/downlink peak data-rate (50 Mbps / 150 Mbps in the considered scenario)
- $O_{PR,U/D}$ is the PDCP-RLC overhead in the uplink/downlink direction (24Mbps / 16Mbps)

Option 6 bandwidth requirement for uplink/donwlink $B_{6,U/D}$ is calculated as:

$$B_{6,U/D} = R_{U/D} + O_{MAC,U/D} \qquad (2)$$

where:

- $O_{MAC,U/D}$ is the MAC-PHY overhead in the uplink/downlink direction (44Mbps / 5Mbps)

Option 7 bandwidth requirement for uplink/downlink $B_{7,U/D}$ is calculated as:

$$B_{7,U/D} = N_{SC} \times N_{sym} \times N_{lay} \times S_{IQ} \times 2 + O_{PHY,U/D} \qquad (3)$$

where:

- N_{SC} is the number of subcarriers (1200 for LTE 20MHz)
- N_{sym} is the number of symbols per millisecond (14 for LTE)
- N_{lay} is the number of MIMO layers
- $S_{R,IQ}$ is the sample bit-width (10 for the uplink and 7 for the downlink)
- $O_{PHY,U/D}$ is the physical layer overhead in uplink/downlink (2.4 Mbps / 9 Mbps)

Option 8 bandwidth requirement for uplink/donwlink $B_{8,U/D}$ is calculated as:

$$B_{8,U/D} = M \times S_{F,IQ} \times 2 \times N_P \qquad (4)$$

where:

- M is the sampling rate (30 Msps)
- $S_{F,IQ}$ is the radio over fiber bit-width (16 in both uplink and downlink)
- N_P is the number of antenna ports (2 for the considered scenario)

III. Software Defined Access Cooperation Scheme

The new Software Defined Access Cooperation Scheme (SD-ACS) proposed in this paper operates according to the workflow shown in Fig. 2.

The SDMN controller is periodically informed by the SDA controller about the present estimate of the bandwidth that the PON can grant to the fronthaul. The bandwidth estimation can be done using different techniques, as detailed in Sec. V. Successively, the SDMN controller runs a *functional split calculation algorithm*, described in Sec. IV, that selects the best suitable FS to be deployed matching the estimated available bandwidth. Then, a fronthaul activation command is sent to the SDA controller specifying the requirements in terms of bandwidth and maximum latency. The SDA controller instructs the OLT to implement the appropriate WBA for the ONUs serving the DU/RUs. This way, the cooperative resource management allows the implementation of lowest layer FS options that would be impossible to support with conventional approaches.

Once the SDMN is informed about the fronthaul bandwidth reservation, it communicates with the CU/DU and the DU/RUs to update the FS. Finally, the fronthaul is activated to support the selected FS.

The requirements of the fronthaul segment have a direct impact on the WBA scheme to be adopted in the PON. DBA schemes are based on report-grant mechanisms and, although more efficient in terms of bandwidth utilization, are not suitable for lower layer FS options requiring less than 250 μ s delay. Fixed Bandwidth Allocation (FBA) represents a feasible solution to guarantee low latency [32], [33]. Moreover, [34], [35] evaluate experimentally the effect on the fronthaul latency that lower-layer FS options can tolerate when the PON-based fronthaul is utilized with DBA enabled/disabled at OLT. The results showed that the end-to-end latency when DBA is enabled is compatible with higher-layer split options (e.g. option 2), and the strict latency requirements of the lower-layer FS options (e.g., Option 6-8) require the activation of FBA at the PON.

To adopt FBA could result in inefficient bandwidth utilization since it allocates a fixed amount of bandwidth even when it is not required by the RU/DUs. In the proposed scheme, the SDA controller implements a SD-WBA that allocates bandwidth and wavelengths for bearing CU/DU-DU/RU communications according to the fronthaul requirements. It allocates fixed bandwidth for low latency split options (6-8) and activates dynamic report-grant based bandwidth allocation for non latency-critical split option 2, as shown in Fig. 3. Moreover, we dimension the amount of bandwidth reserved for the mobile fronthaul by considering maximum traffic load in the mobile network. This assumption guarantees that the fronthaul traffic has sufficient bandwidth to avoid queuing delay at the ONUs.

Nonetheless, when option 2 is selected, some services in the mobile network could still require very low latency that may not be compatible with DBA such as ultra Reliable Low Latency Communications (uRLLC). Since Option 2 enables the exchange of UE-associated signaling [18], we can separate the traffic associated with these services and reserve them a portion of bandwidth through FBA in order to meet latency requirements. However, uRLLC services are expected to be characterized by very small bandwidth requirements, thus the contribution of the uRLLC services can be neglected for the scope of this work.

Wavelength selection for fronthaul traffic is performed by the SDA and OLT based on network operator policies. It is assumed that, since FS option 8 consumes high bandwidth, the RU/DU are assigned to available wavelengths not to saturate the available capacity on each wavelength. For high layer split options wavelength assignment can be performed dynamically. However, the reduction of active wavelengths in the PON is expected to increase the latency experienced by F1 interface due to the increase of queuing time at the ONU [36].

Although the dynamic modification of the implemented FS

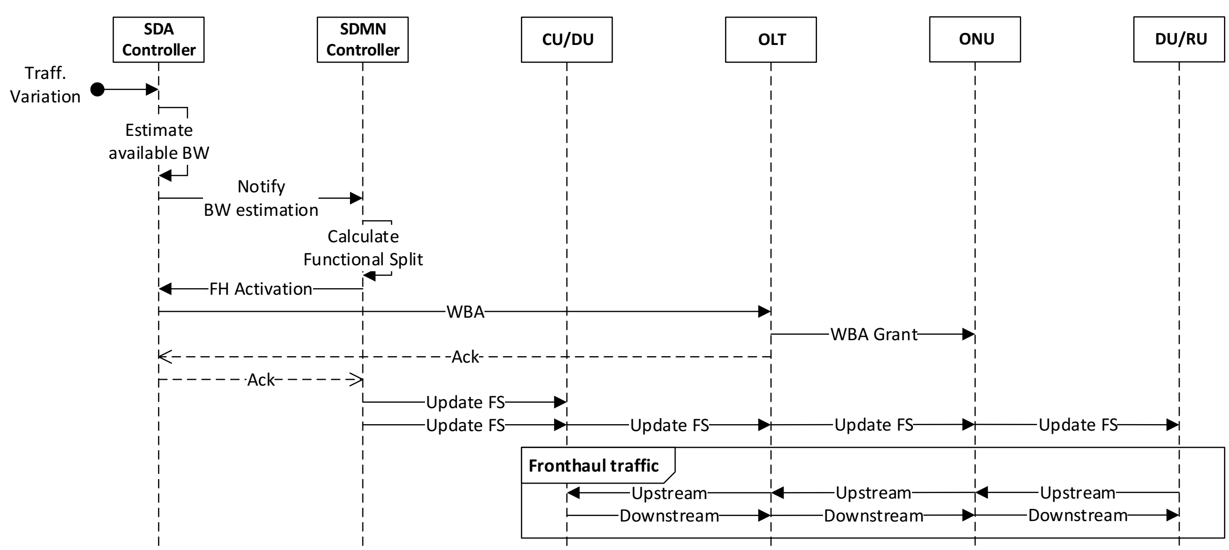


Figure 2: Sequence diagram of the proposed Software Defined Access Cooperation Scheme (SD-ACS).

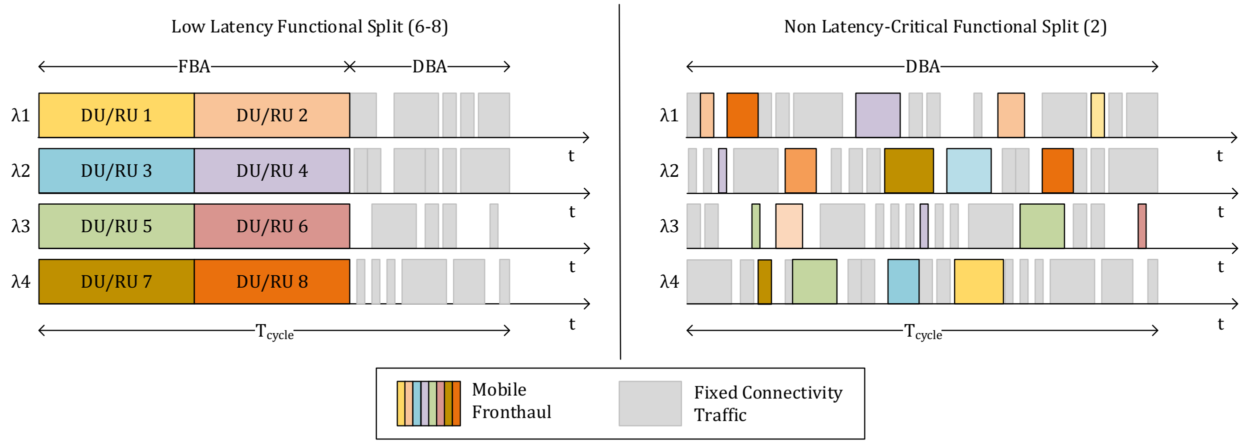


Figure 3: Software Defined Wavelength and Bandwidth Allocation.

between CU/DU and DU/RU highly improves the efficiency of the utilization of both optical and radio resources, it is also a costly operation for the mobile network, because it imposes a service downtime while changing from serving FS to the new FS. Therefore, this operation cannot be performed in arbitrary moments but it has to take place on a periodic basis at specific time intervals established by the mobile operator. We refer to T_{FS} as the time between two consecutive FS decisions, which should be based on a reliable and accurate estimate of the available bandwidth.

Actually, if the available bandwidth is overestimated and the FS with high bandwidth occupancy is implemented, the fixed connectivity may experience a traffic impairment deriving from the excessive bandwidth reservation for the fronthaul. On the other hand, if fixed connectivity traffic bandwidth occupation is overestimated, the FS with lower bandwidth requirements is implemented, leading to a reduction of the achievable performance of the mobile network due to the decentralization of a high number of mobile functionalities at the DU/RUs, thus reducing the centralization advantages. In this work we investigate the suitability of four different techniques for the estimation of available bandwidth, as detailed in Sec. V.

IV. Functional Split Selection Algorithm

This section describes the role of FS selection algorithm in the FFS architecture, because it allows the selection of the best suitable FS to be dynamically implemented in the current period. The selection criterion adopted by the proposed strategy is the maximization of the centralization of functionalities by operating the possible lowest layer FS according to the bandwidth available in the optical PON-based fronthaul.

Indeed, although lower layer split options are more demanding in terms of bandwidth and latency, they enable to leverage centralization advantages in terms of resources allocation, cooperative transmission and QoS management, thus minimizing the overhead and the latency.

The illustrated algorithm runs with periodicity T_{FS} and adapts the FS decision to available bandwidth variations in the PON due to the coexistence with fixed connectivity traffic, i.e. residential FTTH and business connectivity. Let $\hat{B}_{AV}(nT_{FS})$ be the estimated available bandwidth that could be granted to the fronthaul at the instant nT_{FS}, defined as:

$$\hat{B}_{AV}(nT_{FS}) = B_{TOT} - \hat{B}_{FC}(nT_{FS}) \quad (5)$$

where B_{TOT} is the total bandwidth of the TWDM-PON and $\hat{B}_{FC}(nT_{FS})$ is the time-varying estimated bandwidth required by fixed connectivity, that can be calculated with the different approaches presented in Sec. V, each of them having a different impact on the allocated functional split.

Given the available bandwidth estimation $\hat{B}_{AV}$, the proposed algorithm selects the best suitable FS option FS_{max} among the considered options FS_i, defined by 3GPP and listed in Table I. Each option FS_i has specific fronthaul bandwidth requirements B_{R,FS_i}. The total required bandwidth for the fronthaul in the presence of different FS options applied by each DU/RU can be expressed as $B_R^{TOT} = \sum_{j=1}^{N_{DR}} B_{R,FS_j}$. Assuming that the same FS_i is applied to all DU/RUs connected to the same CU/DU, the total bandwidth required at the fronthaul segment connecting the CU/DU to the N_{DR} DU/RUs is $B_{FH_i} = B_{R,FS_i} \times N_{DR}$.

Thus, the problem of individuating the FS with the highest possible centralization level FS_{max} at time nT_{FS} can be formalized as:

$$\begin{aligned} FS_{max} = \max \quad & FS_i \\ \text{subject to} \quad & B_{R,FS_i} \times N_{DR} < \hat{B}_{AV}(nT_{FS}) \\ & i \in \{2,6,7,8\} \end{aligned} \quad (6)$$

This problem can be solved by applying the Algorithm 1, which receives in input from the SDA controller the value of bandwidth available at the PON, $\hat{B}_{AV}$, and iterates over all possible FS_i options. At each iteration it evaluates both the feasibility condition, i.e. if there is enough bandwidth to implement the specified FS_i, and the maximality condition, i.e., if the FS is at the possible lowest layer. As output, at the end of the iteration, the lowest layer FS_{max} that is supportable by the optical network is selected. If the estimated bandwidth request for fixed connectivity services exceeds the available PON capacity the proposed algorithm selects FS option 2 (line 1). In this case in fact the fronthaul traffic is queued at the

Quantity	Notation
Monitoring periodicity [s]	T_M
FS calculation periodicity [s]	T_{FS}
Number of samples per FS interval	N
Duration of a day [s]	T_D
The total bandwidth of the TWDM-PON [Mbps]	B_{TOT}
Bandwidth at time t [Mbps]	$B(t)$
Average bandwidth in $[t_1, t_2]$ [Mbps]	$\overline{B}(t_1, t_2)$
Estimated bandwidth in t [Mbps]	$\hat{B}(t)$
Smoothing factor	α
Smoothing component at time t [Mbps]	$R(t)$
Trend factor	β
Trend component at time t [Mbps]	$G(t)$
Seasonal factor	γ
Season length [samples]	L
Seasonal component at time t [Mbps]	$S(t)$

Table II: Notations

ONU and transmission opportunities are scheduled through the DBA scheme which is assumed to provide allowable latency for option 2.

Algorithm 1 is periodically executed with periodicity T_{FS}. From a complexity viewpoint, the proposed algorithm is an exhaustive search among all the possible FS options, therefore its complexity is $O(h)$ where h is the number of FS options. Since the maximum value of h is equal to 10, the time required to compute the FS is negligible.

Algorithm 1 Functional split calculation algorithm

Input: $N_{DR}, \hat{B}_{AV}(nT_{FS})$
Output: FS_{MAX}
$FS_{MAX} \leftarrow 2$
for all FS Option FS_i **do**
  if $B_{R,FS_i} \times N_{DR} < \hat{B}_{AV}(nT_{FS})$ **and** $FS_i > FS_{MAX}$ **then**
    $FS_{MAX} = FS_i$
  end if
end for
return FS_{MAX}

V. Available Bandwidth Estimation Techniques

The formulation of the four considered forecasting techniques for the accurate estimation of the available bandwidth involves a large number of parameters. For the convenience of the reader, notations are summarized in Table II.

Let's T_M be the monitoring periodicity at which a new value of average bandwidth occupation is available at the SDA controller, and assume that the time between two consecutive FS decisions is an integer multiple of T_M, i.e., $T_{FS} = N \cdot T_M$. The number of bandwidth samples N is decided by the mobile operator when planning FS recalculation periodicity. The considered four estimation strategies apply to data sampled on either a T_{FS} or a T_M basis.

We assume that every T_M the SDA calculates a new bandwidth value $B(iT_M)$, for $i = 1, 2, \ldots \infty$, defined as:

$$B(iT_M) = \overline{B}((i-1)T_M, iT_M), \quad (7)$$

where $\overline{B}(t_1, t_2)$ is the average bandwidth occupancy calculated in the interval $[t_1, t_2]$.

A. Previous Average (PA)

The SDA provides the fixed connectivity bandwidth occupancy estimation $\hat{B}_{FC}(nT_{FS})$ for the time interval $[(n-1)T_{FS}, nT_{FS}]$, as the average estimated bandwidth $\overline{B}((n-1)T_{FS})$ in the previous interval $[(n-2)T_{FS}, (n-1)T_{FS}]$:

$$\hat{B}_{FC}(nT_{FS}) = \overline{B}\left((n-1)\, T_{FS}, nT_{FS}\right). \quad (8)$$

B. Simple Exponential Smoothing (SES)

In this case, the bandwidth occupancy estimation is calculated on a T_{FS} time scale with a smoothing factor $\alpha \in [0, 1]$. The idea behind SES technique is to estimate the bandwidth occupancy for the time interval $[(n-1)T_{FS}, nT_{FS}]$ as the weighted combination of the effective experienced value and the value estimated in the previous time interval $[(n-2)T_{FS}, (n-1)T_{FS}]$, i.e.

$$\begin{aligned} \hat{B}(nT_{FS}) &= \alpha \cdot \overline{B}((n-2)T_{FS}, (n-1)T_{FS}) + \\ &+ (1-\alpha) \cdot \hat{B}((n-1)T_{FS}) \end{aligned} \quad (9)$$

Trivially, it can be noticed that when $\alpha = 1$, $\hat{B}(nT_{FS}) = \overline{B}((n-2)T_{FS}, (n-1)T_{FS})$. Thus, the estimation for the next interval of time is equal to the average effective value of the previous interval. When $\alpha = 0$, the bandwidth estimation is exactly the estimated value for the previous interval, i.e., $\hat{B}(nT_{FS}) = \hat{B}((n-1)T_{FS})$.

C. Last Value (LV)

The bandwidth estimate at the time nT_{FS} is the last value of bandwidth occupancy $B(n \cdot N \cdot T_M)$ at the SDA controller, which is calculated every T_M according to (7):

$$\hat{B}(nT_{FS}) = B(n \cdot N \cdot T_M) \quad (10)$$

where $T_{FS} = N \cdot T_M$.

D. Triple Exponential Smoothing (TES)

It is a forecasting method particularly efficient for time series data characterized by self-similarity [37]. TES modeling takes into account three forecasting factors, namely *smoothing*, *trend* and *seasonality*. We define a *season* as the amount of time that is required to observe the data to repeat the same behaviour. This is appropriate for this work, because our goal is to accommodate fronthaul traffic based on the bandwidth occupation of fixed connectivity, whose behaviour is approximately repeated on a daily basis.

In fact, users tend to use fixed connectivity services with a growing trend during day hours with a peak in the evening, whereas the bandwidth request decreases during night hours and reaches the minimum usage in the night. Therefore, considering one day as a *season* duration, TES method appears particularly appropriate for the forecasting of bandwidth occupancy.

One of the advantages of TES method is that it can produce as output m consecutive estimations for m intervals of time within T_M duration. Since we are interested in forecasting the bandwidth occupation in the time $T_{FS} = N \cdot T_M$ we consider $m = N$.

The estimation of bandwidth occupancy in the interval $[(n-1)T_{FS}, nT_{FS}]$ according to the TES method is given by:

$$\hat{B}(nT_{FS}) = \frac{1}{N} \cdot \sum_{m=1}^{N} \hat{B}((n-1)T_{FS} + m \cdot T_M) \tag{11}$$

where $\hat{B}((n-1)T_{FS}+m \cdot T_M)$ is the m steps ahead forecast, given by

$$\begin{aligned}\hat{B}((n-1)T_{FS} + m \cdot T_M) = \Big[R((n-1)T_{FS} - T_M) \\ +m \cdot G((n-1)T_{FS} - T_M)\Big] \\ \cdot S((n-1)T_{FS} - (L+m) \cdot T_M)\end{aligned} \tag{12}$$

obtained as a generalization of the forecast for *one* step ahead

$$\begin{aligned}\hat{B}((n-1)T_{FS} + T_M) = \Big[R((n-1)T_{FS} - T_M) \\ +G((n-1)T_{FS} - T_M)\Big] \\ \cdot S((n-1)T_{FS} - L \cdot T_M)\end{aligned} \tag{13}$$

The above expressions formulate the forecast as linear combinations of the three above-mentioned forecasting factors, namely the smoothing factor $R(t)$, the trend component $G(t)$ and the seasonal component $S(t)$, defined as:

$$\begin{aligned}R(t) &= \alpha \cdot \frac{B(t)}{S(t - L \cdot T_M)} + (1-\alpha) \cdot (R(t - T_M) + \\ &+ G(t - T_M)\,) \\ G(t) &= \beta \cdot (S(t) - S(t - T_M)) + (1-\beta) \cdot G(t - T_M) \\ S(t) &= \gamma \cdot \frac{B(t)}{R(t)} + (1-\gamma) \cdot S(t - L \cdot T_M)\end{aligned} \tag{14}$$

where $L = T_D/T_M$ is the season length (i.e., the number of samples in a day), T_D is the duration of the day, $\beta \in [0, 1]$ is the trend factor, and $\gamma \in [0, 1]$ is the seasonal factor.

VI. Performance Evaluation and Results

A. Performance metrics

The performance evaluation has the scope of showing the effectiveness of the proposed algorithm and its sensitivity to the different techniques illustrated in Sec. V. Hence, we consider the following performance metrics.

1) Fulfillment index $\eta(kT_M) \in [0, 1]$: We define the *fullfilment index* at the k-th sampling instant as the ratio between the total bandwidth provided to fixed connectivity and the effective total bandwidth required fixed connectivity services at time kT_M, i.e.:

$$\eta(kT_M) = \frac{B_{EFC}(kT_M)}{B_{FC}(kT_M)} \tag{15}$$

where $B_{EFC}(kT_M)$ and $B_{FC}(kT_M)$ are the aggregated bandwidths effectively *granted* to and *required* by fixed connectivity services, respectively.

Therefore, $0 \leq \eta(kT_M) \leq 1$. If the requested bandwidth is totally granted to the fixed connectivity services, we have $B_{EFC}(kT_M) = B_{FC}(kT_M)$ and $\eta(kT_M) = 1$. If $\eta(kT_M) \leq 1$, $B_{EFC}(kT_M) \leq B_{FC}(kT_M)$ and the performance of fixed connectivity is impaired by fronthaul traffic.

2) Impaired traffic: We define *impaired traffic* the amount of fixed connectivity traffic that is transferred at a bit-rate lower than desired, expressed as:

$$I = \frac{1}{8} \sum_{k}^{K} B_I(k \cdot T_M) \cdot T_M, \tag{16}$$

where K sets the width of the temporal observation window as KT_M, and $B_I(kT_M) = B_{EFC}(kT_M) - B_{FC}(kT_M)$ is the bandwidth mismatch between ideal and actually achieved bandwidth.

3) Unallocated supported traffic: We define the *unallocated supported traffic* as:

$$U = \frac{1}{8} \sum_{k}^{K} B_U(k \cdot T_M) \cdot T_M, \tag{17}$$

where $B_U(kT_M)$ is the unallocated bandwidth given by:

$$B_U(kT_M) = B_{TOT} - B_{FC}(kT_M) - B_{FH}(kT_M) \tag{18}$$

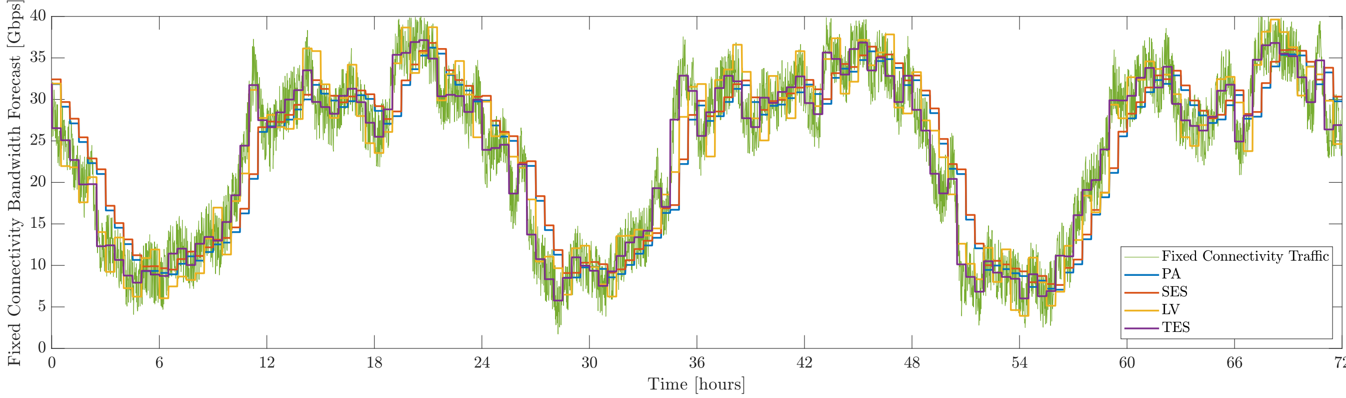
where B_{TOT} is the total bandwidth of the TWDM-PON, $B_{FC}(kT_M)$ is the total bandwidth required by fixed connectivity traffic, and $B_{FH}(kT_M)$ is the fronthaul required bandwidth for the selected FS option.

4) Functional split (FS) uptime: Finally, we consider the *FS uptime* as the percentage of time that each of the selected FS options is running with respect to the total observation time.

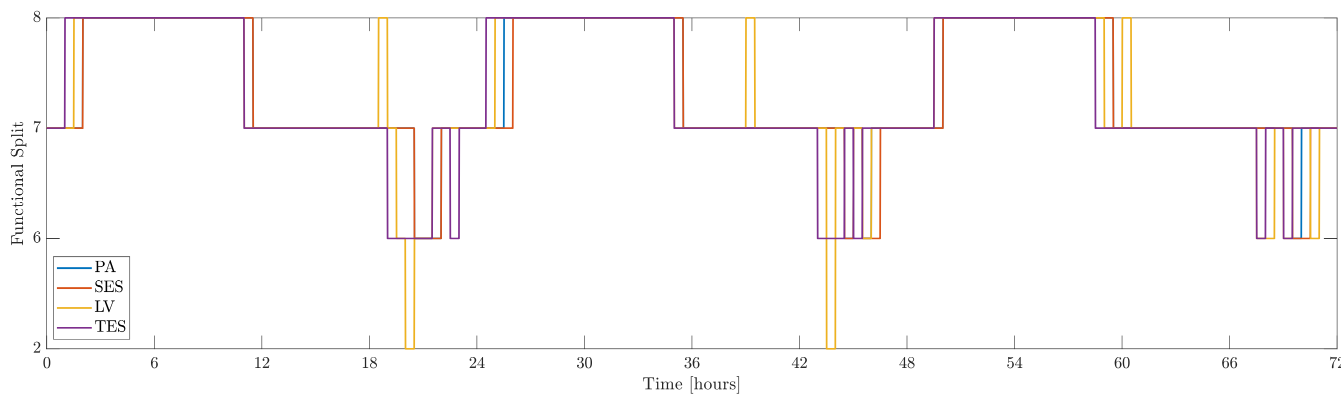
B. Reference scenario

The scenario adopted for the simulation consists of an urban area where 20 DU/RUs are deployed and connected using a PON-based fronthaul infrastructure. The fronthaul is implemented through a TWDM-PON with four wavelength pairs. Wavelengths support 10 Gbps symmetric upstream/downstream data rate. The TWDM-PON is utilized also as access infrastructure for other services characterized by the time-variant upstream traffic load illustrated in Fig. 4a, which is representative of the expected traffic in a NGPON2. To the best of our knowledge, experimental datasets of NG-PON2 traffic associated to fixed connectivity services are not yet available. Hence, we referred to the experimental data measured for a GPON supporting 1.25 Gbps upstream [38], scaled to obtain a 40 Gbps traffic. We also extended the dataset to multiple days by taking the same pattern as a baseline. To differentiate the traffic patterns among different days, we added a stochastic fluctuation to the average recorded pattern available in the literature [38]. We added to every sample a random amplitude uniformly distributed in the interval $[-4; +4]$ Gbps. Finally, we over-sampled the processed dataset with a period of 1 min to match the timescale of the proposed bandwidth allocation strategy, by interpolating with a smoothing function the samples available in [38], which are collected every 15 min.

For the estimation we consider a smoothing factor $\alpha = 0.5$ and a seasonal factor $\gamma = 0.5$. Since we are not interested



(a) Fixed Connectivity Bandwidth Estimation.



(b) Applied functional split option.

Figure 4: Bandwidth estimation Vs. applied FS option (3 days).

in the evaluation of the trend component, we set $\beta = 0$. The numerical evaluation is performed in MATLAB environment.

Regarding fronthaul requirements, we consider the values shown in Table I. Moreover, we assume the SDA controller providing an estimation of the available bandwidth to the SDMN controller every $T_M = 60$ s and FS decision periodicity $T_{FS} = 1800$ s.

Hence, we assume that the TWDM-PON supports these two types of traffic: the traffic due to fixed connectivity services, such as FTTx and business connectivity, and the fronthaul traffic with a load varying according to the FS option currently enforced (refer to Table I). Moreover, the same FS option is considered for all DU/RUs to evaluate the proposed approach. However interested readers can refer to [6], [39] for further deployment scenarios.

C. Analysis of results

The available bandwidth forecast and the corresponding selected FS option are shown in Fig. 4 for a period of three days. Fig. 4a shows the time-variant upstream traffic assumed for the performance evaluation. The bandwidth forecasts calculated with the techniques illustrated in Section V are superimposed to the fixed connectivity services traffic curve and their accuracy is shown. The mismatches between the four estimates of the bandwidth occupancy $\hat{B}_{FC}$ and the actual traffic load lead to different values of available bandwidth for the fronthaul $\hat{B}_{AV}$. Hence, two different forecasting techniques may bring to different FS decisions in the presence of the same amount of real traffic.

Fig. 4b shows the corresponding FSs selected during the day by adopting the proposed cooperative approach.

When the traffic required by the fixed connectivity services in the PON increases, the FS is scaled down to reduce the required bandwidth, preventing any degradation of the QoS offered from the PON to the fixed connectivity services. On the other hand, when the traffic in the PON decreases, more functionalities are centralized and lower layer FS options are implemented. The illustrated scenario shows that FS option 7 is mostly adopted during the largest portion of the day. This represents a good compromise between mobile centralization and bandwidth utilization. The main advantage of these FS options is that traffic aggregation from transmission points can be centralized, i.e. the MAC resides in CU/DU, hence centralized scheduling and joint processing can be applied.

As highlighted in Table I, lower layer FSs require fixed bandwidth reservation to meet delay requirements. This means that when the maximum bandwidth occupancy is reached in the PON, fronthaul traffic bandwidth cannot be allocated dynamically. On the other hand this could imply a degradation of the fixed connectivity traffic flowing in the PON, if $\hat{B}_{AV}$ for the fronthaul is overestimated.

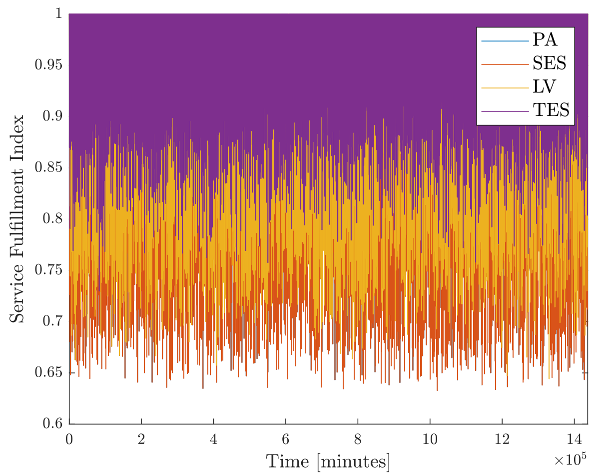


Figure 5: Fulfillment index comparison (1000 days).

Next, we compare the performance of the proposed FFS approach when different estimation techniques are adopted focusing on the performance impact on the fixed connectivity traffic. In Fig. 5 the fulfillment index for an observation interval of 1000 days is illustrated. Results show that TES allows to efficiently address the demand of traffic in the PON. Instead, the other techniques lead to a performance degradation of other supported services, which only a fraction of required bandwidth can be granted to. Indeed, when TES is implemented, fulfillment index η has negative peaks between 0.85 and 1 while for LV it is mostly between 0.85 and 0.75. The worst performance is offered by SES with some peaks of fulfillment index of 0.65. This means that in critical conditions only 65% of the bandwidth required by fixed connectivity services is effectively granted.

This is confirmed by calculating the daily amount of impaired traffic. Fig. 6 shows that the maximum amount of impaired traffic is obtained with simple exponential smoothing, i.e. 4.02 Terabytes (TB). This is reduced to 74% when adopting TES techniques. In fact TES achieves the minimum amount of impaired traffic, that is 1.04 TB.

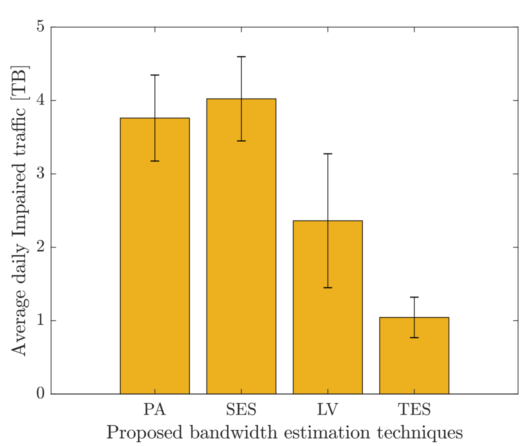


Figure 6: Average daily impaired traffic.

The highest service performance and the reduced amount of impaired traffic can be either explained by two possible reasons:

- overestimation of fixed connectivity bandwidth having as a drawback lower bandwidth for the fronthaul with low centralization of functionalities
- a more accurate forecast that estimates the appropriate amount of bandwidth for the fronthaul between two consecutive scheduling decisions avoiding conflicts of bandwidth allocation and preserving mobile centralization

To investigate this aspect we study the service time of the different split options. Fig. 7 shows the percentage of up-time of the different FS options with respect to a total observation time of 1000 days. PA, LV and TES show similar percentages of time for the different FS options, with small variations. In the case of SES a reduction of Option 8 up-time and an increase of Option 6 up-time are observed compared to the other techniques. This means a lower capability to support the desired centralization of mobile functionalities.

TES better performs in terms of fulfillment index and impaired traffic while keeping the same level of centralization compared to the other techniques. This clarifies that the reduction of fixed connectivity traffic impairment is related to a more accurate bandwidth estimation. On the other hand, PA and LV achieve the same level of centralization but at the same time introduce fixed connectivity traffic impairment. This is due to the less accurate bandwidth estimation that forces the SDMN controller to implement high centralization split options even when bandwidth is not sufficient for the fronthaul, thus resulting in fixed connectivity service impairment.

Finally, the comparison of the different forecasting techniques in terms of the amount of daily additional unallocated traffic that can be supported in the TWDM-PON is shown in Fig. 8. As shown in Sec. VI-A the amount of unallocated traffic is strictly related to the fraction of unallocated bandwidth. From the operator viewpoint the higher is the unallocated additional supported traffic the higher is the possibility to support additional services and increase incomes. Results show that TES offers the same amount of unallocated traffic to other services with a small variation with respect to PA, while achieving better fulfillment index and reduced traffic impairment.

VII. Conclusion

In this work we proposed a converged approach between SDMN and SDA for joint control of flexible functional split and wavelength and bandwidth allocation in TWDM-PON fronthauling. We illustrated a functional split selection strategy based on the available bandwidth in the PON which leverages a software defined wavelength and bandwidth allocation to fulfill bandwidth and latency requirements of the fronthaul. The proposed scheme adapts the bandwidth allocation to the current traffic demand and simultaneously allows the mobile network to take advantage of the highest possible centralization of mobile network functions by leveraging the flexible functional split option adaptively compliant to the current

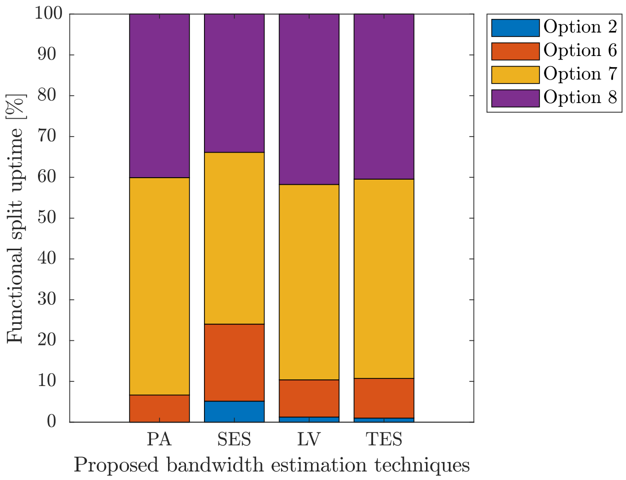


Figure 7: Functional split options uptime comparison.

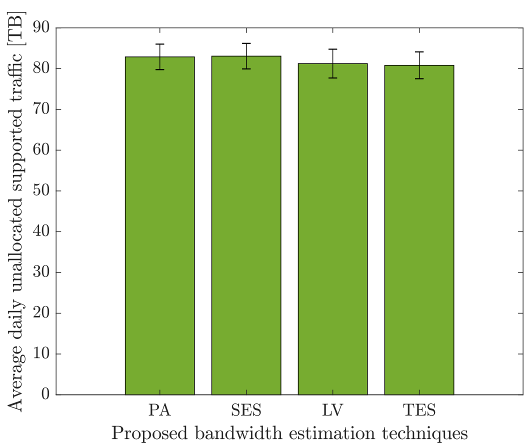


Figure 8: Average daily extra supported traffic.

optical traffic demand. The results show that the proposed converged approach allows to efficiently address the regular traffic demand in the PON, while providing the maximum possible benefit to the mobile network by granting the highest possible degree of centralization of mobile functions permitted by the current traffic conditions in the PON. Techniques for accurate available bandwidth estimation enabling efficient flexible functional split management are investigated. Triple exponential smoothing forecasting technique is shown to minimize traffic impairment while keeping high level of mobile network centralization. Moreover it is shown that TES has comparable performance with other techniques in terms of additionally supported unallocated traffic, leaving space for other services and increasing the incomes of the operator.

Acknowledgment

This work was partially supported by the Italian Government under CIPE resolution no. 135 (December 21, 2012), project INnovating City Planning through Information and Communication Technologies (INCIPICT) and by the EC through the H2020 5G-TRANSFORMER project (Project ID 761536).

References

[1] Deloitte University Press, "A network of networks, how will carriers handle the evolution to 5G?" Report, June 2017.
[2] SDx Central, "2017 NFV Report Series Part 1: Foundations of NFV: NFV Infrastructure and VIM," SDN Central Market Report, 2017.
[3] J. ichi Kani *et al.*, "Flexible Access System Architecture to Support Diverse Requirements and Agile Service Creation," in *Proc. of ECOC*, Sept 2017, pp. 1–3.
[4] 3GPP TR 38.801 , "Study on new radio access technology: Radio access architecture and interfaces," Release 14, Mar. 2017.
[5] L. M. Larsen, A. Checko, and H. L. Christiansen, "A survey of the functional splits proposed for 5G mobile crosshaul networks," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 1, pp. 146–172, 2018.
[6] NGMN - Next Generation Mobile Networks , "NGMN Overview on 5G RAN Functional Decomposition," Tech. Rep., 2018.
[7] Y. Alfadhli, M. Xu, S. Liu, F. Lu, P. Peng, and G. Chang, "Real-Time Demonstration of Adaptive Functional Split in 5G Flexible Mobile Fronthaul Networks," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.
[8] A. de la Oliva, J. A. Hernandez, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, February 2016.
[9] P. Arnold, N. Bayer, J. Belschner, and G. Zimmermann, "5G Radio Access Network Architecture Based on Flexible Functional Control / User Plane Splits," in *Proc. of EuCNC 2017*, Jun 2017.
[10] ITU-T, "Transport network support of IMT-2020/5G," ITU-T Technical Report, Feb. 2018.
[11] 3GPP TR 38.816 , "Study on CU-DU lower layer split for NR," Release 15, Dec. 2017.
[12] A. de la Oliva, J. A. Hernández, D. Larrabeiti, and A. Azcorra, "An overview of the CPRI specification and its application to C-RAN-based LTE scenarios," *IEEE Communications Magazine*, vol. 54, no. 2, pp. 152–159, 2016.
[13] F. Giannone, K. Kondepu, H. Gupta, F. Civerchia, P. Castoldi, A. Franklin, and L. Valcarenghi, "Impact of Virtualization Technologies on Virtualized RAN Midhaul Latency Budget: A Quantitative Experimental Evaluation," *IEEE Communications Letters*, vol. 23, no. 4, pp. 604–607, 2019.
[14] A. Martínez Alba, J. H. Gómez Velásquez, and W. Kellerer, "An adaptive functional split in 5G networks," in *2019 IEEE INFOCOM WKSHPS-3rd Workshop on Flexible and Agile Networks: 5G and Beyond (FlexNets*, 2019.
[15] C.-Y. Chang, N. Nikaein, R. Knopp, T. Spyropoulos, and S. S. Kumar, "FlexCRAN: A Flexible Functional Split Framework over Ethernet Fronthaul in Cloud-RAN," in *Proc. of ICC 2017*, May 2017.
[16] A. Maeder, M. Lalam, A. D. Domenico, E. Pateromichelakis, D. Wbben, J. Bartelt, R. Fritzsche, and P. Rost, "Towards a flexible functional split for cloud-RAN networks," in *2014 European Conference on Networks and Communications (EuCNC)*, June 2014, pp. 1–5.
[17] K. Kondepu, A. Sgambelluri, N. Sambo, F. Giannone, P. Castoldi, and L. Valcarenghi, "Orchestrating lightpath recovery and flexible functional split to preserve virtualized RAN connectivity," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 11, pp. 843–851, Nov 2018.
[18] B. Bertenyi, R. Burbidge, G. Masini, S. Sirotkin, and Y. Gao, "NG radio access network (NG-RAN)," *Journal of ICT Standardization*, vol. 6, no. 1, pp. 59–76, 2018.
[19] A. Garcia-Saavedra, X. P. Costa, D. J. Leith, and G. Iosifidis, "FluidRAN: Optimized vRAN/MEC Orchestration," *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, pp. 2366–2374, 2018.
[20] D. Harutyunyan and R. Riggio, "Flex5G: Flexible Functional Split in 5G Networks," *IEEE Transactions on Network and Service Management*, pp. 1–1, 2018.
[21] 5GPPP Architecture Working Group, "View on 5G Architecture," White Paper, Dec. 2017.
[22] Y. Nakayama, H. Uzawa, D. Hisano, H. Ujikawa, H. Nakamura, J. Terada, and A. Otaka, "Efficient DWBA Algorithm for TWDM-PON with Mobile Fronthaul in 5G Networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017, pp. 1–6.
[23] S. Bidkar, J. Galaro, and T. Pfeiffer, "First Demonstration of an Ultra-Low-Latency Fronthaul Transport Over a Commercial TDM-PON Platform," in *2018 Optical Fiber Communications Conference and Exposition (OFC)*, March 2018, pp. 1–3.
[24] J. P. Elbers *et al.*, "Software-defined access networks," in *Proc. of ECOC*, Sept 2014, pp. 1–3.

[25] T. Chen *et al.*, "Software defined mobile networks: concept, survey, and research directions," *IEEE Communications Magazine*, vol. 53, no. 11, pp. 126–133, November 2015.

[26] T. Tashiro, S. Kuwano, J. Terada, T. Kawamura, N. Tanaka, S. Shigematsu, and N. Yoshimoto, "A novel DBA scheme for TDM-PON based mobile fronthaul," in *OFC 2014*, March 2014, pp. 1–3.

[27] H. Uzawa *et al.*, "Practical Mobile-DBA Scheme Considering Data Arrival Period for 5G Mobile Fronthaul with TDM-PON," in *Proc. of ECOC*, Sept 2017, pp. 1–3.

[28] A. Marotta, K. Kondepu, D. Cassioli, C. Antonelli, L.M. Correia and L. Valcarenghi, "Software Defined 5G Converged Access as a viable Techno-Economic Solution," in *2018 Optical Networking and Communication Conference (OFC)*, Mar. 2018, pp. 1–5.

[29] A. Marotta, D. Cassioli, K. Kondepu, C. Antonelli and L. Valcarenghi, "Efficient Management of Flexible Functional Split through Software Defined 5G Converged Access," in *2018 IEEE International Conference on Communications*, May 2018, pp. 1–6.

[30] M. Jarschel, T. Zinner, T. Hossfeld, P. Tran-Gia, and W. Kellerer, "Interfaces, attributes, and use cases: A compass for SDN," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 210–217, June 2014.

[31] 3GPP TSG RAN WG3, "Transport requirement for CU and DU functional splits options," R3-161813, Aug. 2016.

[32] T. Kobayashi, H. Ou, D. Hisano, T. Shimada, J. Terada, and A. Otaka, "Bandwidth allocation scheme based on simple statistical traffic analysis for TDM-PON based mobile fronthaul," in *2016 Optical Fiber Communications Conference and Exhibition (OFC)*, March 2016, pp. 1–3.

[33] D. Eugui and J. A. Hernández, "Low-Latency Transmission of Fronthaul Traffic over XG(S)-PON with Fixed-Elastic Bandwidth Reservations," in *Optical Fiber Communication Conference (OFC) 2019*. Optical Society of America, 2019, p. W2A.28.

[34] F. Ponzini, K. Kondepu, F. Giannone, P. Castoldi, and L. Valcarenghi, "Optical Access Network Solutions for 5G Fronthaul," in *2018 20th International Conference on Transparent Optical Networks (ICTON)*, July 2018, pp. 1–5.

[35] U.Bucci, A. Marotta, K. Kondepu, C. Antonelli, L. Valcarenghi, and D. Cassioli, "Experimental Demonstration of Flexible TDM-PON Fronthauling," in *Asia Communications and Photonics Conference*. Optical Society of America, 2018.

[36] L. Valcarenghi, K. Kondepu, and P. Castoldi, "Delay fairness in reconfigurable and energy efficient TWDM PON," *Computer Networks*, vol. 110, pp. 223–231, 2016.

[37] P. R. Winters, "Forecasting sales by exponentially weighted moving averages," *Manage. Sci.*, vol. 6, no. 3, pp. 324–342, Apr. 1960.

[38] J. F. Riera, C. Bock, E. Escalona, V. Jungnickel, K. Habel, M. C. Parker, S. Walker, T. Quinlan, V. Marques, and D. Levi, "Software-defined wired-wireless access network convergence: The sodales approach," in *2014 IEEE Globecom Workshops (GC Wkshps)*, Dec 2014, pp. 1522–1527.

[39] P. Chanclou, L. A. Neto, K. Grzybowski, Z. Tayq, F. Saliou, and N. Genay, "Mobile fronthaul architecture and technologies: A RAN equipment assessment [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 1, pp. A1–A7, Jan 2018.