THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Design and Service Provisioning Methods for Optical Networks in 5G and Beyond Scenarios

Maryam Lashgari

Department of Electrical Engineering
Chalmers University of Technology
Gothenburg, Sweden, 2023

Design and Service Provisioning Methods for Optical Networks in 5G and Beyond Scenarios

Maryam Lashgari

ISBN: 978-91-7905-841-8
Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5307
ISSN 0346-718X
This thesis has been prepared using LaTeX.

Department of Electrical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg, Sweden
Phone: +46 (0)31 772 1000
www.chalmers.se

Front cover illustration: General network architecture composed of access, pre-aggregation, aggregation, and core segments.

Printed by Chalmers Reproservice
Gothenburg, Sweden, April 2023

To my family

Abstract

Network operators are deploying 5G while also considering the evolution towards 6G. They consider different enablers and address various challenges. One trend in the 5G deployment is network densification, i.e., deploying many small cell sites close to the users, which need a well-designed transport network (TN). The choice of the TN technology and the location for processing the 5G protocol stack functions are critical to contain capital and operational expenditures. Furthermore, it is crucial to ensure the resiliency of the TN infrastructure in case of a failure in nodes and/or links while the resource efficiency is maximized.

Operators are also interested in 5G networks with flexibility and scalability features. In this context, one main question is where to deploy network functions so that the connectivity and compute resources are utilized efficiently while meeting strict service latency and availability requirements. Off-loading compute resources to large and central data centers (DCs) has some advantages, i.e., better utilization of compute resources at a lower cost. A backup path can be added to address service availability requirements when using compute off-loading strategies. This might impact the service blocking ratio and limit operators' profit. The importance of this trade-off becomes more critical with the emergence of new 6G verticals.

This thesis proposes novel methods to address the issues outlined above. To address the challenge of cost-efficient TN deployment, the thesis introduces a framework to study the total cost of ownership (TCO), latency, and reliability performance of a set of TN architectures for high-layer and low-layer functional split options. The architectural options are fiber- or microwave-based. To address the strict availability requirement, the thesis proposes a resource-efficient protection strategy against single node/link failure of the midhaul segment. The method selects primary and backup DCs for each aggregation node (i.e., nodes to which cell sites are connected) while maximizing the sharing of backup resources. Finally, to address the challenge of resource efficiency while provisioning services, the thesis proposes a backup-enhanced compute off-loading strategy (i.e., resource-efficient provisioning (REP)). REP selects a DC, a connectivity path, and (optionally) a backup path for each service request with the aim of minimizing resource usage while the service latency and availability requirements are met.

Our results of the techno-economic assessment of the TN options reveal that, in some cases, microwave can be a good substitute for fiber technology. Several factors, including the geo-type, functional split option, and the cost of fiber trenching and microwave equipment, influence the effectiveness of the microwave. The considered architectures show similar latency and reliability performance and meet the 5G service requirements. The thesis also shows that a protection strategy based on shared connectivity and compute resources can lead to significant cost savings compared to benchmarks based on dedicated backup resources. Finally, the thesis shows that the proposed backup-enhanced compute off-loading strategy offers advantages in service blocking ratio and profit gain compared to a conventional off-loading approach that does not add a backup path. Benefits are even more evident considering next-generation services, e.g., expected on the market in 3 to 5 years, as the demand for services with stringent latency and availability will increase.

Keywords: 5G and beyond, optical network design, network control and management, cost-efficiency, profit, TCO, functional split, resiliency, latency, availability.

List of Publications

This thesis is based on the following publications:

[A] **Maryam Lashgari**, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti, "Fiber- vs. Microwave-based 5G Transport: a Total Cost of Ownership Analysis". *European Conference on Optical Communication (ECOC), Basel, Switzerland, Sep. 2022.*

[B] **Maryam Lashgari**, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti, "Techno-economics of Fiber vs. Microwave for Mobile Transport Network Deployments (Invited)". *To be published in Journal of Optical Communications and Networking (JOCN), vol. 15, no. 7.*

[C] **Maryam Lashgari**, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti, "Techno-economics of 5G Transport Deployments". *Proc. of Next-Generation Optical Communication: Components, Sub-Systems, and Systems XII, SPIE 12429, San Francisco, California, United States, Jan. 2023.*

[D] **Maryam Lashgari**, Lena Wosinska, and Paolo Monti, "A Shared-Path Shared-Compute Planning Strategy for a Resilient Hybrid C-RAN". *21st International Conference on Transparent Optical Networks (ICTON), Angers, France, July 2019.*

[E] **Maryam Lashgari**, Carlos Natalino, Luis M. Contreras, Lena Wosinska, and Paolo Monti, "Cost Benefits of Centralizing Service Processing in 5G Network Infrastructures". *Asia Communications and Photonics (ACP) Conference, Chengdu, China, Nov. 2019.*

[F] **Maryam Lashgari**, Lena Wosinska, and Paolo Monti, "End-to-End Provisioning of Latency and Availability Constrained 5G Services". *IEEE Communications Letters, vol. 25, no. 6, pp. 1857-1861, June 2021.*

[G] **Maryam Lashgari**, Federico Tonini, Lena Wosinska, Luis M. Contreras, and Paolo Monti, "Next-Generation Service Deployment with Compute Off-Loading: a Profit Analysis Perspective". *Submitted to IEEE Network in Apr. 2023.*

Other publications by the author, not included in this thesis, are:

[H] F. Marzouk, **M. Lashgari**, J. P. Barraca, A. Radwan, L. Wosinska, P. Monti, and J. Rodriguez, "Virtual Networking for Lowering Cost of Ownership". In *Enabling 6G Mobile Networks*, J. Rodriguez, C. Verikoukis, J. S. Vardakas, and N. Passas, Eds. Cham: Springer International Publishing, 2022, pp. 331–369.

[I] **M. Lashgari**, F. Tonini, L. Wosinska, and P. Monti, "Designing and Operating Optical Networks in the 5G and Beyond Era". *To be presented in Advanced Photonic Congress, Busan, South Korea, July 2023.*

[J] M. H. Keshavarz, M. Hadi, **M. Lashgari**, M. R. Pakravan, and P. Monti, "Optimal QoS-Aware Allocation of Virtual Network Resources to Mixed Mobile-Optical Network Slices". *Proc. of IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, Dec. 2021.

Acknowledgments

The research in this thesis would not have been possible without the guidance of my supervisors. First and foremost, I would like to express my sincere gratitude to my supervisor and examiner, Prof. Paolo Monti, for accepting me as his Ph.D. student and for all his guidance, encouragement, and support throughout these years. I would also like to thank my co-supervisor, Prof. Lena Wosinska, for all the fruitful discussions, which helped me better understand my research field. I want to thank my co-supervisor, Dr. Federico Tonini, for the exciting discussions and his invaluable feedback on my work.

I am grateful to Massimiliano Capacchione, Gabriele Rigamonti, and Goran Biscevic for the outstanding collaboration and fruitful discussions and for hosting me on the SIAE premises. I thank Dr. Luis M. Contreras for the great collaborations and for providing an operator's perspective on the project. I want to thank Dr. Carlos Natalino Da Silva for his insightful discussions. Many thanks to Asst. Prof. Marija Furdek Prekratic, for all the support and excellent discussions we had during lunchtime. I am also grateful to Prof. Daniel Kilper for accepting the role of the opponent and to Assoc. Prof. Sofie Verbrugge, Dr. René Bonk, and Prof. Christian Fager for accepting the role of grading committee member for my Ph.D. defense.

I want to thank everyone at the Optical Networks Unit for providing a friendly work environment. I would like to thank Ehsan and Leyla for being great office mates. Living abroad has difficulties, and I am grateful to all my friends who have made Sweden feel like home. I want to thank Björn Östlund, my future manager at Ericsson, for his understanding and flexibility in starting my upcoming journey after the Ph.D. defense.

I want to acknowledge the support of various projects, including MSCA-ITN project 5G STEP FWD funded by the European Union's Horizon 2020 research and EUREKA cluster CELTIC-NEXT project AI-NET-ANIARA, AI-NET-PROTECT, and Smart City Concepts in Curitiba sponsored by VINNOVA.

Last but not least, I would like to express my deepest gratitude to those who have supported me unconditionally during my lifetime and imparted invaluable lessons. Thus, my special thanks belong to my parents and sister, who have been a source of encouragement and inspiration throughout my life, and to my partner Ali, for his love, understanding, and generous support during these years.

Maryam Lashgari,
Göteborg, April 2023

Acronyms

3GPP:	3rd generation partnership project
6-DoF:	6 degrees of freedom
AE:	access edge
AP:	access point
AS:	application server
AVCU:	average compute resources utilization
AVLU:	average connectivity resources utilization
BBU:	baseband unit
C-RAN:	centralized-RAN
CapEx:	capital expenditure
CompU:	compute unit
COTS:	commercial off-the-shelf
CU:	cost unit
D-RAN:	distributed-RAN
DC:	data center
DeMUX:	de-multiplexer
DU:	data unit
eMBB:	enhanced mobile broadband
FA:	fiber aggregation
GPP:	general-purpose processor
H-CRAN:	hybrid cloud radio access network
HLS:	high layer split
IoT:	internet-of-things
ITU:	international telecommunication union

ITU-R:	ITU-radio communication
ITU-T:	ITU-telecommunication
KPI:	key performance indicator
LLS:	low layer split
MA:	metro aggregation
MAC:	medium access control
MC:	macrocell
ME:	metro-core edge
mIoT:	massive Internet-of-things
mMTC:	massive machine type communication
MN:	metro node
MS:	media streaming
MTBF:	mean time between failure
MTTF:	mean time to failure
MTTR:	mean time to repair
MU:	monetary unit
MUX:	multiplexer
MW:	microwave
ND:	networking device
NFV:	network function virtualization
NGMN:	next-generation mobile networks
NPP:	no path protection
NR:	non-reconfigurable
O-RAN:	open RAN
OADM:	optical add-drop multiplexer
OpEx:	operational expenditure
OTA:	over-the-air

PCU:	power consumption unit
PDCP:	packet data convergence protocol
PDN:	passive distribution node
PHY:	physical
PON:	passive optical network
PRS:	preliminary resource sharing
PtP:	point-to-point
QoS:	quality of service
R:	reconfigurable
RAN:	radio access network
RAU:	radio aggregation unit
RCC:	radio cloud center
RD:	resource duplication
REP:	resource-efficient provisioning
RF:	radio frequency
RIRS:	reconfiguration and improved resource sharing
RL:	reinforcement learning
RLC:	radio link control
ROADM:	reconfigurable optical add-drop multiplexer
RRC:	radio resource control
RRU:	remote radio unit
SC:	small cell
SDN:	software-defined networking
SLA:	service level agreement
SPSCP:	shared-path shared-compute planning
TCO:	total cost of ownership
TN:	transport network

TU:	time unit
Tx/Rx:	transceiver
UE:	user equipment
UP:	user plane
UPF:	user plane function
URLLC:	ultra-reliable low latency communications
URLLC-S:	URLLC-latency-sensitive
URLLC-T:	URLLC-latency-tolerant
V2X:	vehicle-to-X
VR:	virtual reality
WDM:	wavelength division multiplexing

Contents

Part I

Overview

CHAPTER 1

Introduction

5G networks and beyond are expected to revolutionize communication by providing ubiquitous and fast mobile connectivity and delivering services with stringent requirements. 5G technology provides various service categories [1], [2] including 1) enhanced mobile broadband (eMBB): needs a large amount of capacity to support high-speed data transfer, 2) ultra-reliable low latency communications (URLLC): demands extremely low latency (i.e., the time delay measured over the path from the user to the core network) and high reliability for delivering error-free data packets within a bounded time, and 3) massive machine type communication (mMTC): allows a large number of devices to connect to the network, with each device sending small amounts of data.

The service demands are getting more stringent with the advent of 6G and new services such as extended reality, digital twins, holographic-type, and haptic communications [3], [4]. In order to meet the strict service requirements of 5G and prepare for the evolution towards 6G scenarios, operators need to consider various aspects and enablers.

Network densification is an enabler for delivering the promised high-capacity, low-latency, and high-reliability communication [5]. Network densi-

fication refers to deploying new cell sites close to the user. The cell sites must be connected to the core network through a transport network (TN). To fully enable the benefits of network densification, a cost-efficient solution for TN deployment that can meet the service requirements is of utmost importance.

For the TN deployment, operators need to decide on the radio access network (RAN) architecture [6]. They may choose a fully distributed architecture where the 5G protocol stack functions are processed in the cell site. In this architecture, the segment between the cell site and the core network is called backhaul. Another option is to deploy the 5G protocol stack functions in central data centers (DCs). In this option, the segment between the cell site and the DC is called fronthaul. The third option is to split the protocol stack functions between the cell site, a distributed unit, and a central unit. The segment between cell sites and distributed units in a TN is known as fronthaul, while the segment between distributed units and central units is referred to as midhaul. Lastly, the segment between central units and the mobile core network is backhaul [7], [8]. Each segment is comprised of various nodes and links. It is also possible to have only one splitting point instead of two and deploy 5G protocol stack functions in the cell site and a DC.

Eight splitting point options are defined in 5G protocol stack layers [6]. A typical functional split option is high layer split (HLS), where most of the functions in the protocol stack are deployed at the cell site [6], [9]. The latency and capacity requirements over the TN are not stringent for HLS. Another common functional split option is low layer split (LLS) [10], [11], where most of the protocol stack functions are deployed in centralized DCs. This option requires a TN that can offer ultra-low latency and very high capacity. The advantage of LLS is that coordinated processing among cell sites and load balancing is possible.

For the TN deployment, operators decide on a functional split option (e.g., HLS or LLS). They also need to select a technology for the TN, which is a critical decision affecting the total cost of ownership (TCO), latency, and network connection availability. The operators aim to minimize the capital expenditure (CapEx) and operational expenditure (OpEx) while meeting service demands.

Fiber and microwave are two major technologies typically used for TN deployment, bringing different pros and cons in terms of capacity, cost, and ease of installation [12], [13]. Fibers can provide very high capacity, but the

deployment is expensive and time-consuming. On the contrary, deploying microwave devices on existing towers is relatively fast and inexpensive, but their capacity and reach are more limited than fibers [12], [14]. In general, fiber connections are considered to be more reliable and offer better connection availability than microwave links as they are not affected by environmental conditions [12], [14].

Operators also need to decide on the reconfigurability of the TN architecture, which will impact the capability to adapt to changing traffic demand. A reconfigurable TN can support traffic growth over time without requiring equipment upgrades to a certain extent. However, reconfigurable equipment is typically more expensive than non-reconfigurable alternatives, affecting the initial deployment cost. In contrast, with reconfigurable TN, the increased cost is deferred to future upgrades.

Operators must understand the implications of a particular choice of functional split option, technology, and reconfigurability levels on the TCO, latency, and connection availability performance to determine the most promising solution for the TN deployment.

However, even with a suitable TN deployment and careful planning, failures may occur in various components and locations, causing service disruptions for the users in the affected areas. The failures in the DCs or the nodes/links of the midhaul segment can affect many users [15]. Therefore, operators must adopt a failure recovery approach in the midhaul network to minimize the negative impacts on users. To achieve this, they must develop strategies to design a resilient midhaul architecture while maximizing resource efficiency.

A conventional approach to ensure the survivability of services in the event of failures is to provide dedicated backup resources [16]. In this way, the backup resources can substitute the primary resources when they fail. Thus, the services can be delivered despite the failure. However, providing backup resources can increase the cost, and they stay idle most of the time, assuming that failures in the network only happen sometimes. To improve the resource efficiency in a survivable network, backup resources can be shared wherever possible and depending on the failure scenario [16]. However, implementing methods that maximize sharing among backup resources is more complex than conventional methods, which only provide dedicated backup resources. Therefore, it is crucial to investigate the potential reduction in resource usage that can be achieved by leveraging resource sharing.

In 5G networks and beyond, a diverse range of services with varying requirements is expected to be provisioned. Deploying these services on a shared infrastructure allows for cost-effective network design and enables efficient use of network resources [17]. Consequently, operators are looking into a common infrastructure that can provide a programmable, multi-purpose, flexible, and scalable platform supporting such services. Software-defined networking (SDN) and network function virtualization (NFV) are two enablers for these capabilities [18], [19]. SDN is an architecture that decouples the control plane (i.e., the part of a network that carries control and management traffic) from the data plane (i.e., the part of the network that carries user traffic) [17], [19]. NFV helps operators to virtualize their resources and instantiate virtualized network functions on commercial off-the-shelf (COTS) servers and general-purpose processors (GPPs) instead of implementing functions in specialized hardware [20]. One of the major challenges in NFV is to decide where to deploy virtualized network functions so that the connectivity and compute resources are used efficiently, and the service requirements are met [19]. SDN and NFV can greatly assist in service provisioning and optimizing infrastructure resource utilization.

5G services such as URLLC require extremely low latency and high-reliability performance [21]. Accordingly, these services are restricted to being deployed as close to the user as possible in DCs at edge nodes, which offer better latency. However, edge DCs have limited compute resources. It can be, therefore, an advantage to off-load compute resources by placing services that do not require very low latency in large and centralized DCs. The advantages are many-fold such as: 1) the amount of compute resources in large DCs is abundant, 2) the cost of operations in large and central DCs is lower than edge DCs, thanks to the economy of scale [22], [23] (i.e., cost per unit of resources decreases as the scale of deployment increases), and 3) connectivity resources in higher-tier TN allow the multiplexing of traffic into fewer channels [24]. However, the service availability requirements might be violated when services are deployed in the centralized DCs as these DCs are usually far from end users. One solution is to add a backup path to improve the availability and facilitate compute off-loading for deploying services in cost-efficient central DCs. Utilizing extra connectivity resources on the backup path might increase the cost of service deployment. Thus, the impact of these additional resources on the cost and the profitability of compute off-loading strategies

must be evaluated. Assessing this trade-off is crucial for 6G scenarios, where the service latency and availability requirements are even more stringent.

This thesis proposes innovative methods to address the challenges mentioned above regarding TN design and service provisioning. In the following, we describe the specific research questions and the contributions of this thesis.

1.1 Research Questions

As already explained, the operators need to decide on the technology and reconfigurability capabilities of their TN architecture based on the given functional split option. They need to evaluate TCO, latency, and connection availability to choose the most convenient option. Moreover, a resource-efficient scheme that can guarantee the survivability of services in case of failure in the network is required. Finally, the resource efficiency and profitability evaluation of compute off-loading strategies for service provisioning are important.

This thesis proposes cost-efficient design and service provisioning methods in 5G networks and beyond. In particular, the following questions are addressed in this thesis:

- What are the TCO, latency, and connection availability implications of fiber- vs. microwave-based TN deployment using HLS and LLS? What are the impacts of having reconfigurability features on these performance metrics?
- How to maximize sharing of backup resources for resilient service delivery considering a single failure scenario either in DCs or the nodes/links of the midhaul segment? How much can the total resource usage be reduced compared to conventional resilient network design strategies?
- What are the advantages of compute off-loading strategies in a dynamic service provisioning scenario considering service latency and availability requirements? In particular, what are the impacts of backup-enhanced compute off-loading strategies on resource utilization, service rejection ratio, and operator's profit? How do the advantages vary in future scenarios (e.g., three years and five years ahead)?

In the following subsection, we summarize the contributions of this thesis addressing the research questions above.

1.2 Thesis Contributions

The contributions of this thesis can be divided into three parts, each one addressing a different research question: a) evaluation of TCO, latency, and connection availability of fiber- vs. microwave-based TN for HLS and LLS options, b) proposing a resource-efficient strategy for a resilient midhaul network design, and c) proposing a backup-enhanced compute off-loading strategy for dynamic service provisioning with latency and availability constraints such that resource-efficiency is maximized. The summary of each contribution is presented in the following.

Fiber vs. Microwave for Transport Network Deployments: a Techno-economic Analysis

This study aims to assess several deployment options for TN when densifying the wireless network by adding new cell sites. A comprehensive framework is proposed to evaluate the TCO, latency, and connection availability of a set of TN architectures.

We consider the HLS and LLS options for the architectures based on fiber or microwave. We also investigate the impact of using reconfigurable equipment. We consider URLLC and eMBB services and their requirements to evaluate the latency and connection availability performance of studied architectures.

Regarding the scenarios and the network modeling, we leverage data from real deployments of a large mobile network operator in a city in South America. Three geo-types are investigated with different area sizes, number of cell sites, distances between sites, and capacity requirements. This contribution is made in collaboration with a system vendor (i.e., SIAE Microelettronica), who also provided real data on network dimensions and components cost.

To examine the broad applicability of our conclusion, we do a sensitivity analysis on the cost of fiber trenching and microwave equipment. These cost values can differ among operators and countries, depending on the labor cost and the negotiated price of microwave equipment and fiber trenching [25]–[27].

The results in **Paper A** show the TCO, including a sensitivity analysis of fiber and microwave-based architectures for HLS option. The network dimensioning is the initial stage deployment of 5G (i.e., the number of cell sites is not large enough to provide 5G coverage everywhere). Results in **Paper B** indicate the TCO, latency, and connection availability performance of fiber-

and microwave-based architectures for both HLS and LLS options in a mature stage deployment of 5G (i.e., the number of cell sites are sufficient to realize the full advantages of 5G). It also considers the reconfigurability capabilities for HLS architectures. The conclusions of **Paper A** and **Paper B** are that the TCO gains of microwave vs. fiber vary depending on different functional split options (HLS or LLS), geo-types, fiber trenching cost, and negotiated cost of microwave equipment. The impact of these factors is more evident in dense urban geo-type than in urban and sub-urban. In fact, in dense urban, microwave-based architectures have comparable TCO to fiber-based ones for the LLS option. In dense urban geo-type, the average link length is relatively short. Thus, the cost of fiber trenching (in fiber-based architectures) is balanced by the high cost of microwave devices (in microwave-based architecture). In other geo-types (i.e., urban and sub-urban), using microwave leads to higher TCO gains compared to fiber (due to the long average link length and higher cost of fiber deployment).

Moreover, the considered architectures have similar latency and connection availability performance, and they can meet the requirements of eMBB and URLLC services. However, in LLS, for a service class with stringent latency requirements (i.e., 0.025 [ms]), a small percentage of sites cannot fully meet the requirement when using microwave-based architecture. This is due to the latency of microwave links, which do not allow using multiple microwave hops to provide a very stringent latency requirement. For this extreme scenario, a hybrid fiber-microwave architecture is considered in **Paper C**. The results show that the hybrid architecture can contain the TCO compared to fiber-only-based architectures by using microwave where possible. At the same time, the hybrid architecture can resolve the strict latency requirement issue (i.e., 0.025 [ms]) by connecting the troublesome sites by fiber.

Resilient Network Design by a Shared-Path Shared-Compute Strategy

In case of a link or node failure in the TN, a large number of users will be affected at the same time. Therefore, the TN must be resilient against failures to avoid user service interruption. **Paper D** addresses this problem by proposing a network recovery strategy referred to as shared-path shared-compute planning (SPSCP), which guarantees the survivability of services from a single failure of either nodes/links of midhaul segment or cloud DCs.

We assume several cell sites are connected to an aggregation node, which, in turn, must be connected to a cloud DC (to forward traffic of its connected cell sites to the cloud DC). The SPSCP strategy assigns primary and backup servers (located in two different cloud DCs) to each aggregation node. The aggregation node is connected to its related primary and backup DCs via two node-disjoint paths. SPSCP tries to maximize sharing of backup connectivity and compute resources among aggregation nodes.

The simulation results demonstrate the cost-efficiency of SPSCP expressed by the cost saving of up to 28% compared to a benchmark method based on dedicated backup resources. Besides, the cost saving of 23% is obtained compared to another method which first assigns primary and backup DCs to aggregation nodes and then reconsiders/changes pairing between aggregation nodes and backup DCs to improve sharing.

Resource-efficient Service Deployment with Latency and Availability Constraints

5G and beyond services have strict latency, availability, compute, and connectivity requirements. The infrastructure resources are limited and must be utilized efficiently for service provisioning. As explained earlier, large DCs have abundant compute resources, and deploying services on those DCs has many advantages.

This work aims to maximize the resource efficiency of service provisioning while meeting stringent service requirements. For this purpose, the thesis presents a backup-enhanced compute off-loading strategy where services are deployed in large and centralized DCs as much as possible. This contribution is made in collaboration with an operator to make realistic assumptions on network dimensions and service types in future networks.

The relatively simple scenario presented in **Paper E** was a first step to show the potential benefits of compute off-loading. The results indicate that compute off-loading can lead to 74% savings of the total communication infrastructure cost.

The analysis of compute off-loading is then extended to a dynamic service provisioning scenario in **Paper F**. We propose a backup-enhanced compute off-loading method, referred to as resource-efficient provisioning (REP), which also considers service latency and availability demands. For each service request, REP selects a DC (to deploy the required compute resources by the

service), a connectivity path (to connect the cell site and selected DC), and (optionally) a backup connectivity path. The selection is based on a metric that maximizes connectivity and compute resources efficiency. In **Paper F**, two types of services with different latency and availability requirements are considered. Results indicate that using REP as a service provisioning strategy can result in up to four orders of magnitude reduction of service blocking ratio compared to a conventional method as a benchmark. The benchmark does not add a backup path in service provisioning, which limits its capability for compute off-loading. The service blocking ratio gains of REP reach two orders of magnitude in the case of provisioning service requests with more relaxed latency and availability demands.

Paper G of this thesis presents a profitability analysis of REP in a dynamic service provisioning scenario with multiple types of service requests within the network. This analysis offers a guideline to operators considering compute off-loading as a provisioning strategy for next-generation services, which have strict requirements regarding latency, availability, compute, and connectivity resources. Three scenarios were created to simulate current, short-term (i.e., three years from now), and long-term (i.e., five years from now) traffic predictions. Each scenario features a different composition of existing services in next-generation networks. Results indicate the profit gain and lower service blocking ratio of REP compared to a conventional approach (i.e., the benchmark strategy of **Paper F**). The advantages mentioned above are particularly evident when considering short- and long-term traffic predictions, as operators must accommodate more demanding next-generation services.

1.3 Thesis Outline

The thesis is organized as follows:

- Chapter 2 defines essential concepts and background information to follow the rest of the thesis, including 5G/6G service requirements, functional split options, network survivability aspects, and latency and availability modeling.
- Chapter 3 presents several TN architectures based on fiber and microwave technologies. The TCO, latency, and availability models are discussed. Moreover, the TCO, latency, and availability performance of architectures under exam are evaluated, and results are presented.

- The resilient midhaul network design is discussed in Chapter 4. The performance evaluation of the proposed resilient design strategy is compared against benchmark methods.
- Chapter 5 describes the backup-enhanced compute off-loading strategy. The network architecture, latency, availability, and profit models are presented. Moreover, the performance of REP in terms of blocking ratio, resource utilization, and profit is illustrated.
- Chapter 6 provides a summary of all the papers included in the thesis.
- The thesis concludes in Chapter 7 with final remarks and potential areas for future research.

CHAPTER 2

Background Information and Concepts

Operators are seeking resource-efficient approaches for deploying 5G networks that meet increasing traffic demand and diverse requirements of various services. To achieve this goal, it is essential to investigate different deployment options, model the service requirements, and explore methods to meet the service demands while minimizing overall costs.

In this chapter, we explain various concepts and provide background information to introduce the research areas addressed in the thesis. This includes a brief explanation of 5G and beyond service demands, followed by an example of a network architecture. Then, various RAN architectures and standardization efforts for the evolution from distributed RAN to centralized RAN, and finally, functional split options are described. Finally, network protection schemes, availability, and latency modeling are discussed, which are needed to model performance metrics.

2.1 5G and 6G Service Requirements

5G has promised to provide high-speed, low-latency, and high-reliability communications. The industry stakeholders have defined various use cases and

services for 5G [20]. Based on that, the international telecommunication union-radiocommunication (ITU-R) has characterized three service categories shown in the vertices of the triangle in Fig. 2.1 [1]:

1. eMBB: This service category requires mobile broadband connectivity and high capacity for data-intensive applications, such as content delivery networks and video streaming.
2. mMTC: This category aims to connect many devices to the mobile network, each with low data volume. Smart homes and industrial internet-of-things (IoT) are examples of services within this category.
3. URLLC: This service category requires high network reliability and low latency. This can refer to mission-critical applications, autonomous driving, and object tracking.

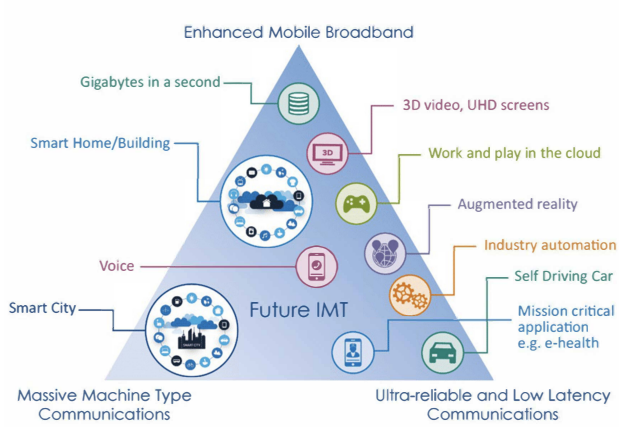


Figure 2.1: 5G use cases defined by ITU [1].

The emerging services, such as holographic-type and haptic communications, ubiquitous intelligence, tactile internet, and digital twins [3], [4] are defined for the next-generation mobile networks referred to as 6G networks. These services are more demanding than the ones specified in 5G. To support such services, the network performance must be drastically improved.

6G is an advanced technology that integrates communication, storage, control, sensing, and compute capabilities [28]. The key performance indicators (KPIs) of 6G compared to 5G are illustrated in Fig. 2.2. It is shown that

different performance indicators need to be 10 to 100 times better than in 5G.

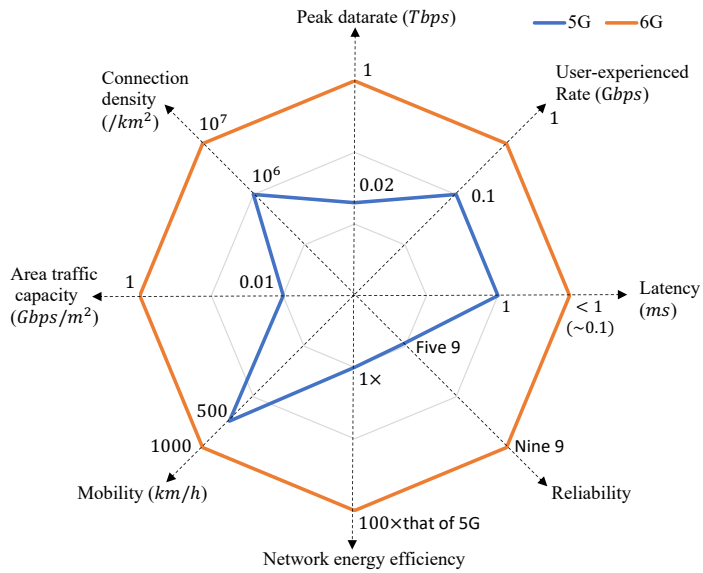


Figure 2.2: Key performance indicators (KPIs) comparison between 5G and 6G [4], [28]–[31].

The KPIs are briefly introduced as follows:

- **Peak data rate** of 6G is expected to be higher than 1 [Tbps] to support applications such as holographic communication and tactile internet. This rate is at least 50 times larger than that of 5G networks [31].
- **User-experienced data rate** of 6G is 1 [Gbps] which is 10 times that of 5G. For some scenarios, the expected data rate is even higher, i.e., in the order of 10 [Gbps] [30].
- **Latency** requirement in 6G is less than 1 [ms] (around 100 [μs]), which extremely improves the quality of experience for the users [28].
- **Reliability** is essential for emergency services and other applications such as cooperative autonomous driving and industrial automation. The reliability target in 6G is 99.9999999%, while it was 99.999% in 5G [28], [31].
- **Energy efficiency** requirement of 6G is 10-100 times higher than 5G to

decrease power consumption and provide a cost-efficient network. Therefore, 6G can reduce carbon emissions and move towards a green network [4].

- **Mobility** demand is 1000 [km/h] to support services such as high-speed trains, while it was 500 [km/h] in 5G [4].
- **Area traffic capacity** is expected to be 1 [Gbps/m^2], which is needed for indoor hot spots, while it was 10 [Mbps/m^2] for 5G [4], [30].
- **Connection density** is envisioned to be 10^7 per km^2 which is 10 times higher than 5G. Thus, 6G can provide mMTC use case for many devices [4].

Out of these aspects and KPIs, we mainly focus on latency and reliability performance metrics in this thesis. A high-reliability performance can be achieved by benefiting from network protection schemes, which will be discussed in Section 2.4. The latency modeling will be explained in Section 2.6.

To move beyond 5G, many research activities on 6G have been initiated in both industry and academia to meet future demands of information and communications technology (ICT) [4]. In July 2018, the international telecommunication union-telecommunication (ITU-T) standardization sector formed a focus group named *Technologies for Network 2030*. The goal of this group was to study the networks for the year 2030 and beyond. 3rd generation partnership project (3GPP) is planning to start studying 6G around 2025 so that the first commercial roll-out of 6G can be accomplished by 2030. Other consortia have launched similar activities, e.g., the next-generation mobile networks (NGMN) and European Commission [4], [32].

To provide 5G and beyond services, a suitable TN needs to be designed to meet the demanding service requirements. An example of a TN architecture is introduced in the next section.

2.2 Network Architecture

As explained in Chapter 1, mobile network densification enables providing 5G services that need high capacity, low latency, and highly reliable communication. Similar to the existing sites in the network, the newly deployed cell sites need to be connected to the mobile core. The network segment that connects cell sites and mobile core is referred to as the transport network (TN).

Figure 2.3 shows a general network architecture that includes different ag-

gregation layers, nodes, links, base stations, users, and DCs. According to open RAN (O-RAN) consortium, the TN may be composed of several segments, i.e., access, pre-aggregation, aggregation, and long-haul core [10], [33]. The access segment often has a tree or ring structure with a diameter of 10-20 [km], covering various parts of a city. The pre-aggregation and aggregation segments may have different topologies (e.g., ring or mesh) and aggregate several access segments. Each aggregation ring can have a diameter of around 40-80 [km] [34]. The long-haul core segment usually has a mesh architecture.

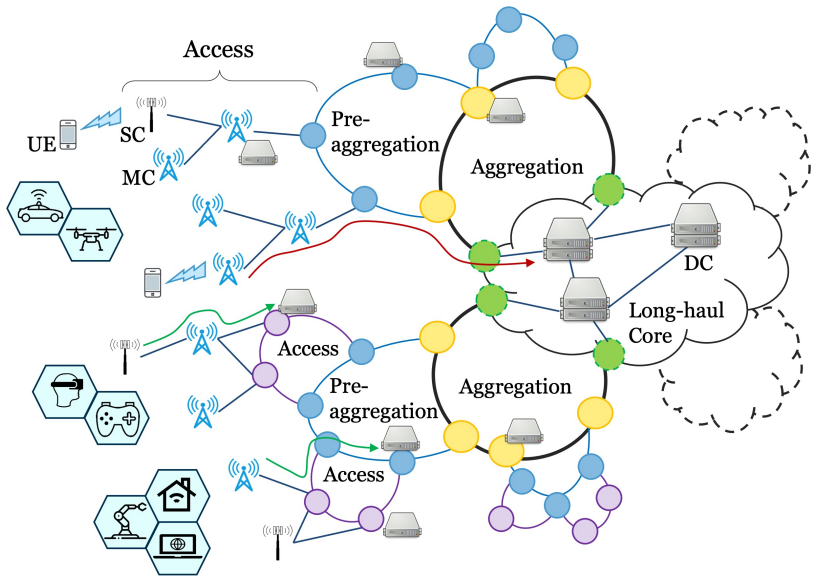


Figure 2.3: General network architecture composed of access, pre-aggregation, aggregation, and long-haul core segments. The users' equipment (UEs) are connected to small cells (SCs) and macrocells (MCs) in the access segment, while data centers (DCs) are located in various segments.

Various components must be placed in the network, i.e., used for connectivity (e.g., fiber cables, microwave links, routers, switches, etc.) and compute purposes (e.g., servers and GPP). Indeed, DCs and GPPs need to be deployed in the network to handle large volumes of data, host services and applications, process the 5G protocol stack functions, and support mobile core functionalities. The DCs can be deployed either in proximity to the user (e.g., in the

access segment) or at a distance from them (e.g., in the aggregation or long-haul core segment).

Using a scalable TN solution that meets the service requirements at the lowest deployment cost is essential from the network operators' perspective. To minimize the cost, operators need to decide on many aspects, such as selecting the appropriate technology for the TN (e.g., links based on fiber, microwave, and satellite interconnects), deciding on RAN architecture (which will be explained in Section 2.3), and adopting energy efficiency measures (e.g., using energy-efficient devices, minimizing power usage).

There are two major options for the TN deployment: 1) greenfield, where everything is deployed from scratch, and 2) brownfield, which refers to the upgrading and installation of new equipment and components on top of an existing infrastructure [22]. Operators can select one of these approaches depending on their needs and whether they have access to existing infrastructure in the area.

Once the TN is deployed, services need to be provided. Dynamic service provisioning refers to the process of selecting a DC, allocating compute resources on that DC, and choosing a connectivity path between the user and the DC such that specific service requirements are met. The DC is required to run the application server or the 5G protocol stack functions. This task is challenging as the deployed compute and connectivity resources are limited and must be used efficiently. The service provisioning process also involves setting up and managing required network elements (e.g., routers, switches, microwave devices, optical equipment, etc.), followed by monitoring the quality of service (QoS).

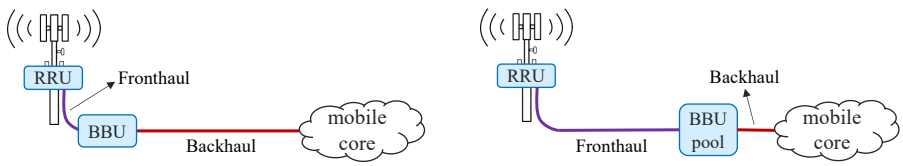
One of the aspects that operators need to consider for a cost-efficient TN deployment is the choice of RAN architecture. Different options for the RAN architecture are explained in the next section.

2.3 RAN Architecture and Functional Split

Operators have various alternatives for the radio access network (RAN) architecture. The conventional option is a distributed-RAN (D-RAN), where all radio and baseband processing functions are deployed at the cell site. The segment between the cell site and mobile core network is called backhaul (shown in Fig. 2.4a) [10]. However, D-RAN has inefficient resource utilization and

high operational costs. Moreover, in order to benefit from inter-cell coordination, a new architecture was needed [8].

4G network introduced the concept of centralized-RAN (C-RAN), which is showen in Fig. 2.4b. In the conventional C-RAN architecture, only the radio frequency (RF) functions are deployed in the remote radio unit (RRU) (located in the cell site tower). All the baseband processing functions are deployed in a central location, referred to as baseband unit (BBU) pool [10]. The network segment between the RRU and the BBU pool is referred to as fronthaul, and the segment between the BBU pool and the mobile core is backhaul.



(a) Distributed-RAN (D-RAN) architecture.

(b) Centralized-RAN (C-RAN) architecture.

Figure 2.4: Distributed-RAN and Centralized-RAN architectures [8]. The remote radio unit (RRU) is in the cell site, while the baseband unit (BBU) is at the cell site or in a centralized location.

The baseband processing centralization brought many benefits, such as 1) minimizing the cost of site rental, operation, and maintenance, 2) sharing compute resources in the BBU pool and achieving multiplexing gain on BBUs, 3) enhancing network performance and better coordination among cell sites to cancel interference, 4) improved resource utilization and cost/energy savings, and 5) faster handover between cell sites that use the same BBU pool [7], [8]. However, the main drawbacks of C-RAN are strict latency and high capacity demands over the fronthaul. This is even more critical with the huge traffic volume in 5G networks [7].

One solution to solve the extreme requirements on the fronthaul is to deploy more baseband processing functions at the cell site and do signal processing before transmitting to the BBU pool. Accordingly, the concept of the functional split was introduced. The functional split option determines the 5G protocol stack functions deployed in the cell site and the functions centralized in a DC [7].

The set of possible splitting points is defined by 3GPP [6], [7] and shown in Fig. 2.5. There are eight functional split options: Option 1 (between radio resource control (RRC) and packet data convergence protocol (PDCP)), Option 2 (between PDCP and radio link control (RLC)), Option 3 (intra RLC), Option 4 (between RLC and medium access control (MAC)), Option 5 (intra MAC), Option 6 (between MAC and physical (PHY)), Option 7 (intra PHY), and Option 8 (between PHY and RF).

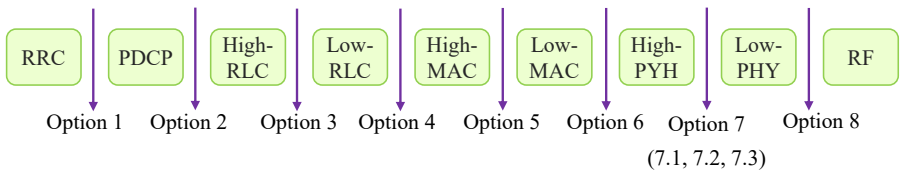


Figure 2.5: Functional split options proposed by 3GPP [6], [7].

If the splitting point is within the high layers of the protocol stack (i.e., the left side of Fig. 2.5), it is referred to as high layer split (HLS). The HLS option is similar to the D-RAN architecture with a backhaul type of traffic. Otherwise, the functional split option is referred to as the low layer split (LLS) (i.e., the splitting point is on the right side of Fig. 2.5). The LLS option has similar features as the C-RAN with a fronthaul type of traffic.

Using the LLS option leads to better coordination and load balancing among cell sites, efficient resource utilization by sharing resources at the centralized location, and a reduced amount of equipment at the cell site [7], [22]. The disadvantages of the LLS option are that it requires high capacity and strict latency over the fronthaul. On the other hand, the HLS option has more relaxed latency and capacity requirements than LLS. However, HLS has limited resource sharing that can lead to less efficient resource utilization and higher operational costs than LLS.

3GPP mainly focuses on split option 2 (i.e., split between PDCP and RLC layers), which is an HLS option, and recognizes it as the most straightforward option to standardize [6], [9]. On the other hand, O-RAN alliance focuses on split option 7.2x (i.e., a split within the PHY layer of 5G protocol stack), which is an LLS option [10], [11].

Another option for the RAN is an x-haul architecture (shown in Fig. 2.6), in which we can leverage the features of both centralized and distributed ar-

chitectures. In an x-haul architecture, two splitting points in the 5G protocol stack can be selected among eight possible split options in Fig. 2.5 [6]. The upper layers of the protocol stack are deployed in a central unit, whereas the middle layers are deployed in a distributed unit. The rest of the functions in lower layers of the protocol stack are deployed in a remote radio unit (RRU) located in the cell site [6]. The segment between the RRU and the distributed unit is referred to as fronthaul, between the distributed unit and central unit is midhaul, and from the central unit to the mobile core network is backhaul.

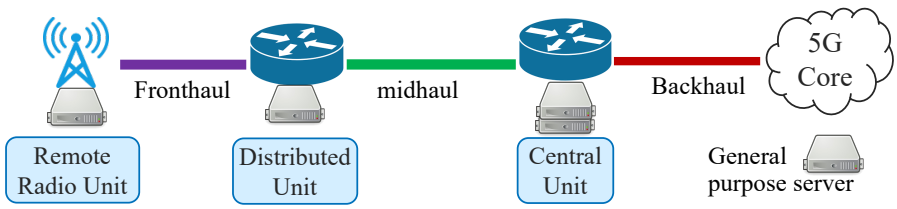


Figure 2.6: RAN architecture [10].

Some of the TN architectures can support the flexible functional split. In these architectures, the 5G protocol stack functions can be split and moved flexibly between the central and distributed units. Selecting the appropriate functional split option depends on various factors, e.g., 1) splitting the functions should be feasible according to the performance of TN in terms of capacity and latency, and 2) the user density and network load, which impact the required level of RAN coordination, and consequently, the selected functional split option [6].

In addition to selecting the appropriate RAN architecture, network survivability is essential for enhancing reliability performance when deploying a TN. We discuss network survivability aspects in the next section.

2.4 Failures and Network Survivability

Network failures can occur for many reasons, e.g., equipment or software malfunctions, fiber cuts, environmental factors, or sabotage (malicious attacks). Therefore, ensuring network resiliency is crucial, which can be achieved by providing backup resources.

Protection and *restoration* are two main techniques to provide network

resiliency. The *protection* strategy is pro-active, i.e., the potential failures are considered before they happen, for which the backup resources are pre-reserved during the designing/provisioning phase. Thus, backup resources can be rapidly available in considered failure scenarios, resulting in fast recovery and 100% survivability. However, this level of survivability is achieved at the cost of large resource consumption as the pre-reserved backup resources cannot be used for other purposes [35]–[37]. Moreover, the network is not resilient against unaccounted failures.

Another option to provide resiliency is *restoration*, which is a reactive approach, and the recovery solutions are computed on-the-fly in the event of a failure. These strategies do not pre-reserve backup resources but instead search for the backup option after a failure has been detected. Thus, restoration's potential to retrieve from unaccounted incidents is advantageous. However, the required time for the signaling in restoration makes it a relatively slow mechanism compared to the protection strategy. Moreover, finding a recovery solution cannot be guaranteed (i.e., cannot ensure 100% survivability) as the free resources might not be available when needed [35]–[37].

According to the service level agreement (SLA) between the infrastructure provider and its customer, the allowed time to recover from a failure can be very stringent (e.g., less than 50 [ms]) or more relaxed in a best-effort manner, where no backup resources are considered for protection [16].

There are two types of protection schemes: *dedicated* and *shared* protection. Two main types of dedicated protection are 1+1 and 1:1. To understand their difference better, let us focus on protecting the primary path. In the 1+1 scheme, the backup path is active simultaneously with the primary path, and the destination chooses the received data between the two paths. Therefore, the recovery from a failure in the primary path is fast. Moreover, if any component on the backup path fails, it can be detected promptly since the backup path must be active simultaneously with the primary path. However, the 1+1 protection scheme requires more components at the connection endpoints to support two active paths (primary and backup path) [16]. In the 1:1 approach, the failure must first be detected, and then, the backup path will be activated. Thus, recovering from a failure takes slightly longer than 1+1. However, by using 1:1 protection, the available spare capacity can be used to carry low-priority traffic. This traffic can be discarded when the capacity is needed for the failure recovery [16].

Shared protection is a resiliency method that requires fewer backup resources than dedicated protection, as it allows using backup resources for multiple primary paths. By assuming a single failure scenario, the failure in the node or link is repaired before another failure occurs in the network. Therefore, the only condition to share resources among multiple primary paths is that they are node-disjoint (i.e., they do not have any common nodes/links). The disadvantage of shared protection is that it can add higher complexity to network design and operation than dedicated protection. Moreover, the recovery is slower than dedicated protection as some switches may need reconfiguration to establish the protection path [16].

The following section discusses how the connection availability is modeled to evaluate the network performance.

2.5 Availability Modeling

Availability and *reliability* are two terms used extensively and sometimes interchangeably to evaluate the performance in a network.

Reliability is the probability that a system will operate without any disruption for a predefined period of time [37].

Availability is the asymptotic probability that a system will be found in the operating state at an arbitrary time. The computation of system availability can be done statistically, which depends on the frequency of failures and the repair rate of the network components used in the system [37], [38]. According to this definition, the availability of an element i is calculated as [39]–[41]:

$$A_i = \frac{MTTF_i}{MTBF_i} = \frac{MTTF_i}{MTTR_i + MTTF_i}. \tag{2.1}$$

In Eq. 2.1, $MTTF_i$ is the mean time to failure of an element i, and during this time, the element is up and running. $MTTR_i$ is the mean time to repair of element i, which depends on the time to diagnose a fault, decide on the procedure to resolve the fault, and send the field team to fix it. $MTBF_i$ is the mean time between failure of element i which is the summation of mean time to failure and mean time to repair.

The availability of a system can be calculated as a function of the availability of its components/elements depending on if they are connected in parallel or

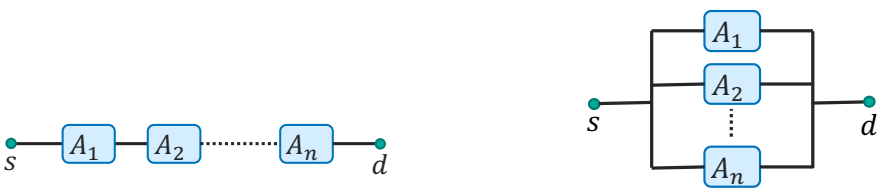
series from the reliability point of view [42]. Suppose the components are connected in series (from the reliability point of view) between a source s and a destination d (as Fig. 2.7a). In that case, the connection is available if all components are available at the same time. Accordingly, the availability between s and d in Fig. 2.7a is calculated as:

$$A_{total} = A_1 \times A_2 \times \cdots \times A_n \tag{2.2}$$

In contrast, the components may be connected in parallel between source s and destination d (as in Fig. 2.7b). In this case, if at least one of the components is up and running, the connection between s and d would be available. Thus, the availability between s and d is calculated as:

$$A_{total} = 1 - ((1 - A_1) \times (1 - A_2) \times \cdots \times (1 - A_n)) = 1 - \prod_{i=1}^{n} UA_i \tag{2.3}$$

where $UA_i = 1 - A_i$. Similarly, the unavailability between s and d is calculated as $UA_{total} = 1 - A_{total}$.



(a) Series connection. **(b)** Parallel connection.

Figure 2.7: Series and parallel composition of components [42].

2.6 Latency Modeling

The user plane (UP) latency is the time that it takes to successfully deliver a packet from the user to the user plane function (UPF) (i.e., a component in 5G architecture responsible for packet inspection, data forwarding, and QoS management). The latency contributors are shown in Fig. 2.8 and UP latency is defined as [10], [43], [44]:

$$l_{\mathrm{UP}} = l_{\mathrm{UE}} + l_{\mathrm{OTA}} + l_{\mathrm{BBU}} + l_{\mathrm{prop}}^{\mathrm{TN}} + l_{\mathrm{switching}} \times n_{\mathrm{switch}}, \tag{2.4}$$

$$l_{\mathrm{BBU}} = l_{\mathrm{RRU}} + l_{\mathrm{DU}} + l_{\mathrm{CU}}, \qquad l_{\mathrm{prop}}^{\mathrm{TN}} = \sum_{i=1}^{p} l_i \tag{2.5}$$

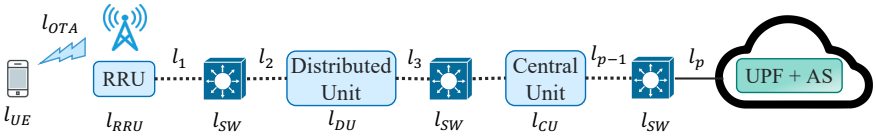


Figure 2.8: The contributors to the latency model.

l_{UE}, l_{OTA}, and l_{BBU} are the latency contributions due to the user equipment (UE), the over-the-air (OTA) propagation between the user and the cell site, and the BBU processing, respectively. l_{BBU} is the summation of latency contribution due to processing 5G protocol stack functions in RRU, distributed unit, and the central unit. $l_{\mathrm{prop}}^{\mathrm{TN}}$ refers to the propagation delay over the TN, i.e., the sum of propagation delay over all links (i.e., l_1 to l_p). $l_{\mathrm{switching}}$ is the switching latency introduced by each switch (devices that perform layer 3 processing), and n_{switch} is the number of switching devices on the path.

To improve the latency performance, a few approaches can be considered: 1) reduce the distance between the UE and UPF, 2) use technology for the physical medium with better latency performance, 3) use switches with shorter switching times, and 4) use more powerful devices for baseband processing. The existing technology affects some of these solutions, and we may be unable to significantly reduce the total latency. Shortening the distance between the UE and UPF is a straightforward strategy to reduce latency.

When we look at the existing works in the literature, there is no unique definition for the latency requirement. ITU-T recommends the UP latency requirements for eMBB and URLLC to be 4 [ms] and 1 [ms], respectively [43]. Moreover, their main focus is on the HLS option. Thus, we can consider UP latency as the target KPI for HLS option. On the other hand, O-RAN alliance considers LLS (with split option 7.2x) and defines various classes with different latency requirements on the fronthaul interface. Some of these classes are

referred to as High25, High75, High100, and High500 with latency requirements of 25[μs], 75 [μs], 100 [μs], and 500 [μs], respectively [10]. Therefore, we can consider these classes for the latency requirement in the LLS option.

The introduced models provide a basis for our analysis in the following chapters.

CHAPTER 3

Fiber vs. Microwave for Transport Network Deployments: a Techno-economic Analysis

5G services have stringent requirements regarding high capacity, low latency, and high reliability. Examples of such services are eMBB and URLLC [2]. Operators must consider different aspects, such as the functional split option, technology choice, and reconfigurability levels for the design and deployment of a transport network (TN) so that the deployment cost is low while the service requirements are met.

This chapter presents various TN architectures based on fiber and microwave for high layer split (HLS) and low layer split (LLS) options. Then, the total cost of ownership (TCO), latency, and connection availability performance of different deployment options are evaluated. Accordingly, operators can select an appropriate TN solution for a given urban area (which has a specific cell site density and average link length).

3.1 Literature Review

Mobile network operators constantly seek innovative deployment and technology options to minimize the cost of their network deployments. The work in [45] proposed an open-source framework for TCO and capacity evaluation of several 5G deployment options using fiber for the backhaul. The authors considered different infrastructure-sharing strategies between two or more mobile network operators. They evaluated the related cost savings, which can be up to 50% saving in the case of infrastructure-sharing among multi-operators. The work in [27] presented a framework to analyze the TCO and economic viability of 5G networks when using fiber and microwave technologies for the TN deployment. The results in this work reveal that to contain the economic benefits of heterogeneous network deployment, choosing a suitable technology for the TN is crucial. Moreover, the work showed that selecting a low TCO solution might not necessarily lead to an increased profit. Instead, the timing of investment in a long-term project can significantly impact the overall profitability. In [46], the authors explored a combination of wired and wireless backhaul technologies for network deployment. They assessed the TCO across three distinct geo-type scenarios with varying user densities. Their findings indicate that a microwave-based approach is more cost-effective than a fiber-based alternative. The authors in [47] assessed the total cost of ownership of a network deployment using fiber (point-to-point (PtP) and passive optical network (PON) options) or wireless technologies. Their work focused on the backhaul segment of a fixed-wireless access use case. Their results show that wireless backhaul is a more economical option when the cost of fiber is high. Otherwise, PtP fiber is a more cost-effective choice.

The studies discussed up to this point have concentrated on the backhaul or HLS option. However, it is also crucial to examine the cost-performance of a TN that employs an LLS option, should an operator choose to implement this approach. Using LLS presents more significant challenges in meeting capacity and latency requirements, necessitating more expensive equipment. The works in [48] and [49] evaluated the TCO of a 5G TN considering various technologies (e.g., wireless, optical) and functional split options (i.e., LLS and HLS). The results presented in [48] show the TCO values associated with four 5G verticals, where the authors assumed different infrastructures depending on the use case. The results indicate that, for one use case, the TCO of the HLS option is higher than that of LLS, while for another use case, the TCO

trend is reversed (i.e., TCO of LLS option is larger). The findings outlined in [49] show that using LLS and a hybrid composition of both wireless and optical technologies can yield a cost-effective solution for TN deployment.

The works in [27], [45]–[47] assessed the TCO performance of fiber- and/or microwave-based architectures only for the backhaul or HLS option. The TCO of fiber- and microwave-based architectures is evaluated in [48], [49] by considering different functional split options (i.e., HLS and LLS). These works do not assess the reconfigurability, latency, and reliability performance of their examined architectures. However, these performance metrics must be evaluated as a given TN option may not be applicable to 5G networks or result in overall performance degradation.

3.2 Network Architectures

This section presents the considered architectures for the TN. Figure 3.1 shows a general TN topology composed of access, pre-aggregation, and aggregation segments as specified in [33].

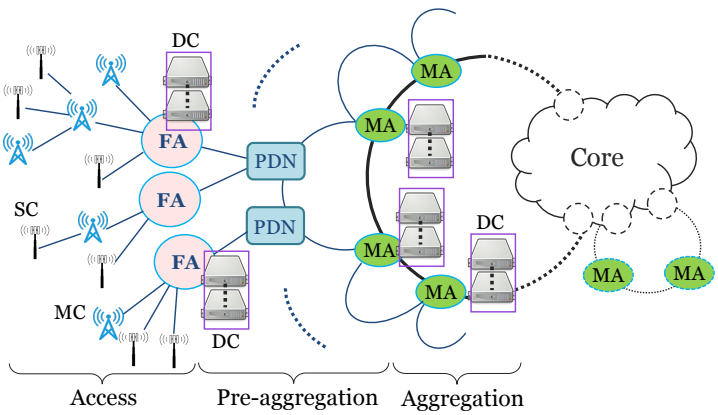


Figure 3.1: Network architecture. Macrocells (MCs) and small cells (SCs) are in the access segment connected to fiber aggregation (FA) nodes. Data centers (DCs) are deployed in FA or metro aggregation (MA) nodes. Reprinted with permission from **Paper B** ©Optica Publishing Group.

The aggregation segment is connected to the core network and consists

of metro aggregation (MA) nodes interconnected by fiber rings. The pre-aggregation rings are connected to the MA nodes and the access segment (through fiber aggregation (FA) nodes), and they include passive distribution nodes (PDNs). In the access segment, the cell sites (macrocells (MCs) and small cells (SCs)) are connected to the FA nodes using either fiber or microwave. The DCs can be deployed in the MA or FA nodes and are used for processing 5G protocol stack functions. We assume that the operator is non-incumbent, meaning it needs to pay the cost of using fiber on the pre-aggregation ring to an incumbent operator. However, to connect MCs and SCs in the access segment, the non-incumbent operator should deploy fibers under the ground. We consider the HLS and LLS options and present the fiber- and microwave-based architectures.

High layer split architectures

For the HLS option, we assume a functional split option 2 [6], [9], [50]. As explained in Section 2.3, in split option 2, most of the processing functions of the 5G protocol stack (i.e., PHY, MAC, RLC) are implemented at the cell site, except a few (i.e., RRC and PDCP), which are deployed in the central DCs [6], [50].

We consider two options for the pre-aggregation segment shown in Fig. 3.2: non-reconfigurable (NR) and reconfigurable (R). The R option provides more flexibility in the pre-aggregation segment than the NR architecture. This advantage is achieved by using reconfigurable optical add-drop multiplexers (ROADMs), while the NR option uses passive optical add-drop multiplexers (OADMs).

The R and NR pre-aggregation segments are connected to an access segment, which uses either fiber or microwave. In HLS, we consider three architectures for the access segment shown in Fig. 3.3. The first one is PtP fiber connections between the cell site and FA ($H^F_{NR,1}$ and $H^F_{R,1}$), the second one is PON-like architecture ($H^F_{NR,2}$ and $H^F_{R,2}$) [51], and the third one is based on microwave links in a tree structure (H^{MW}_{NR} and H^{MW}_{R}). The networking devices (NDs) perform link/network layer processing and traffic grooming. The NDs at the cell sites have similar functionalities (traffic aggregation) as the cell site routers defined in O-RAN [33].

In HLS architectures, the path from the FA to the MA is protected using a backup path over the pre-aggregation ring. This path is protected because the

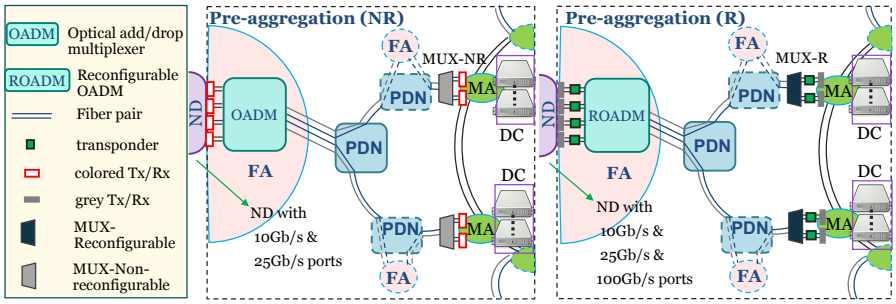


Figure 3.2: Pre-aggregation architecture: non-reconfigurable (NR) and reconfigurable (R). Reprinted with permission from **Paper B** ©Optica Publishing Group.

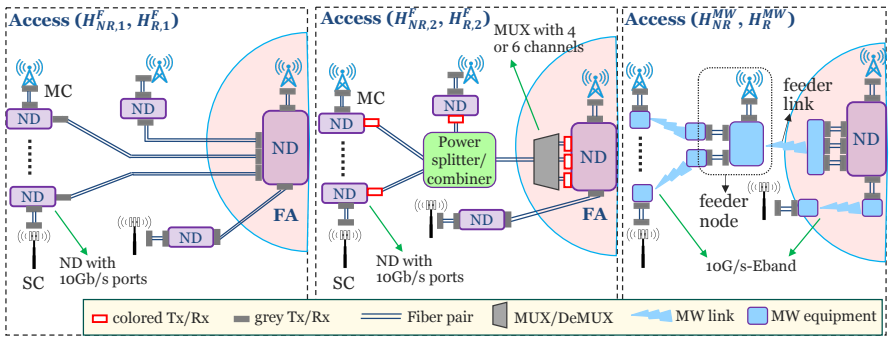


Figure 3.3: Architecture for access segment when considering high layer split (HLS) option. Reprinted with permission from **Paper B** ©Optica Publishing Group.

traffic of many cell sites and users passes through it, and its failure significantly impacts network performance. The switching between the primary and backup path is performed by NDs deployed at the FA. However, the path from the cell site to FA is not protected as a limited number of cell sites are connected through this path. Thus, the impact of a failure on this path is not as high as a failure in the pre-aggregation ring.

Low layer split architectures

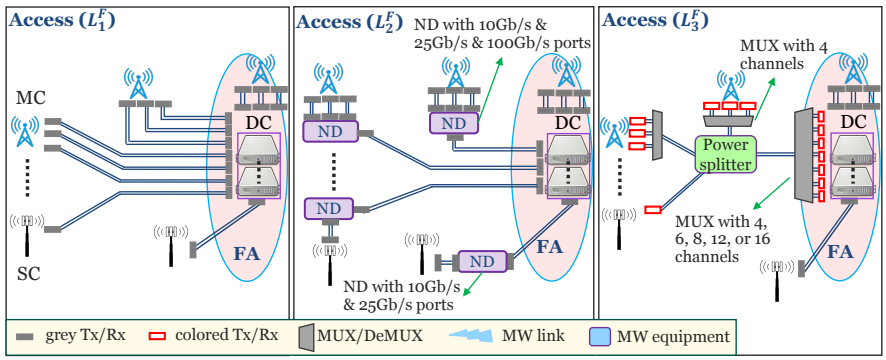
b We assume the split option 7.2x, recommended for LLS by O-RAN. In contrast to the HLS option, in LLS, most of the processing of the 5G protocol stack functions are deployed in the DC located at an FA node (i.e., high-PHY, MAC, RLC, PDCP, and RRC). A small portion of the processing is done at the cell site (i.e., low-PHY and RF) [10], [11].

For LLS, we consider three fiber-based architectures, L_1^F, L_2^F, and L_3^F, illustrated in Fig. 3.4a, as well as one microwave-based (L^{MW}) and one hybrid fiber-microwave-based ($L^{MW,F}$) architectures shown in Fig. 3.4b. L_1^F is based on PtP fiber links, L_2^F aggregates traffic at the cell sites by an ND and uses PtP links for connection to the FA, and L_3^F is a PON-like architecture. The L^{MW} architecture is based on microwave links and uses microwave and mmWave band devices. L^{MW} has two options depending on the required capacity over the feeder link. Due to technological limitations, aggregating all the traffic over the feeder link using a microwave connection may not always be possible. Suppose a single microwave or mmWave device can meet the required capacity over the feeder link. In that case, we use microwave technology for the feeder link (the option on the left side in the illustration of L^{MW}). Otherwise, we use fiber for the feeder link (the option on the right side in the illustration of L^{MW}). Finally, the $L^{MW,F}$ architecture is a hybrid fiber-microwave option. $L^{MW,F}$ can be necessary due to the stringent latency requirements of LLS and the related limitations of multi-hop microwave links. Accordingly, only sites that are one hop away and directly connected to the FA can use the microwave and mmWave band devices. The remaining sites on the feeder are connected by fiber to offer the latency that meets the requirement of the LLS option.

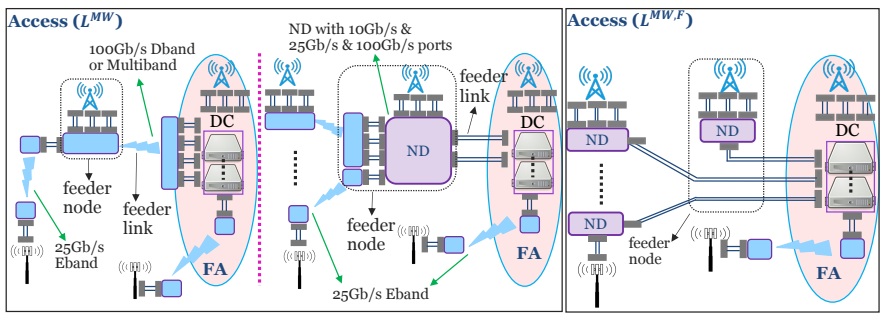
The described architectures use various components with different capabilities and costs. Accordingly, to select a suitable TN option, the cost, latency, and availability performance of architectures should be weighed against each other. In the following, we describe the models to evaluate the performance of the considered architectures.

3.3 Modeling System Performance

This section presents the TCO, latency, and availability performance models to evaluate the considered architectures.



(a) Fiber-based architectures.



(b) Microwave-based and hybrid fiber-microwave architectures.

Figure 3.4: Architectures for access segment when considering low layer split (LLS) option. **Paper C** and reprinted with permission from **Paper B** ©Optica Publishing Group.

TCO model

Figure 3.5 presents the main contributors to the TCO calculation. The CapEx includes the cost associated with the access segment, pre-aggregation segment, and compute resources. The cost of access segment consists of the Ac_{opt} (i.e., optical equipment, router ports, and NDs), Ac_{MW} (i.e., microwave equipment), and Ac_{fib} (i.e., fiber deployment). The cost of pre-aggregation includes PA_{fib} (i.e., fiber deployment from the FA to the PDN and using fiber on pre-aggregation rings) and PA_{opt} (i.e., router ports and optical components in the pre-aggregation segment). The cost of the compute equipment includes

the price of the GPP servers used for processing 5G protocol stack functions. The contributing factors to each cost category are shown in Fig. 3.5, and the details about how each is calculated are explained in **Paper B**.

The OpEx during one year of operation is assumed to be proportional to the CapEx [45], and the energy cost (Fig. 3.5). η_1 and η_2 are the multiplication factors applied to the related CapEx values. We assume the microwave and fiber deployments are easier to maintain over time. Thus, their operational costs are lower than that of optical equipment and servers (i.e., $\eta_2 < \eta_1$). The spectrum license is the licensing fee to use microwave frequencies. The energy cost is related to the energy consumption of all the active components in the considered architectures. The transceivers (Tx/Rxs) energy consumption is neglected as it is negligible compared to other equipment.

Latency characterization

For evaluating the latency performance of the considered architectures, we take into account two service categories of 5G, i.e., eMBB and URLLC [1], [2].

As explained in Chapter 2, different consortia focus on various functional split options. We consider UP latency requirements as the target to assess the performance of the HLS architectures [43]. The model to calculate the UP latency is defined in Eq. 2.4 in Section 2.6, and here, we refer to it as l_{HLS}. For LLS, we consider the transport latency over the fronthaul [10] (l_{LLS}) which comprises some of the elements in Eq. 2.4 (i.e., $l^{\text{TN}}_{\text{prop}} + l_{\text{switching}} \times n_{\text{switch}}$).

In this work, $l_{\text{switching}}$ and n_{switch} (defined in Eq. 2.4) represent the latency introduced by each ND and the number of NDs from the cell site to DC, respectively, where baseband processing takes place. $l^{\text{TN}}_{\text{prop}}$ refers to propagation delay over the fiber and microwave links.

We define the latency performance metric as **Paper B**:

$$P(L) = \frac{\text{Num. of Sites with } l \leq L}{\text{Total Num. of Sites}} \times 100, \tag{3.1}$$

where l is l_{HLS} or l_{LLS} for HLS or LLS options, respectively. $P(L)$ shows the percentage of cell sites that can meet a latency requirement value of L.

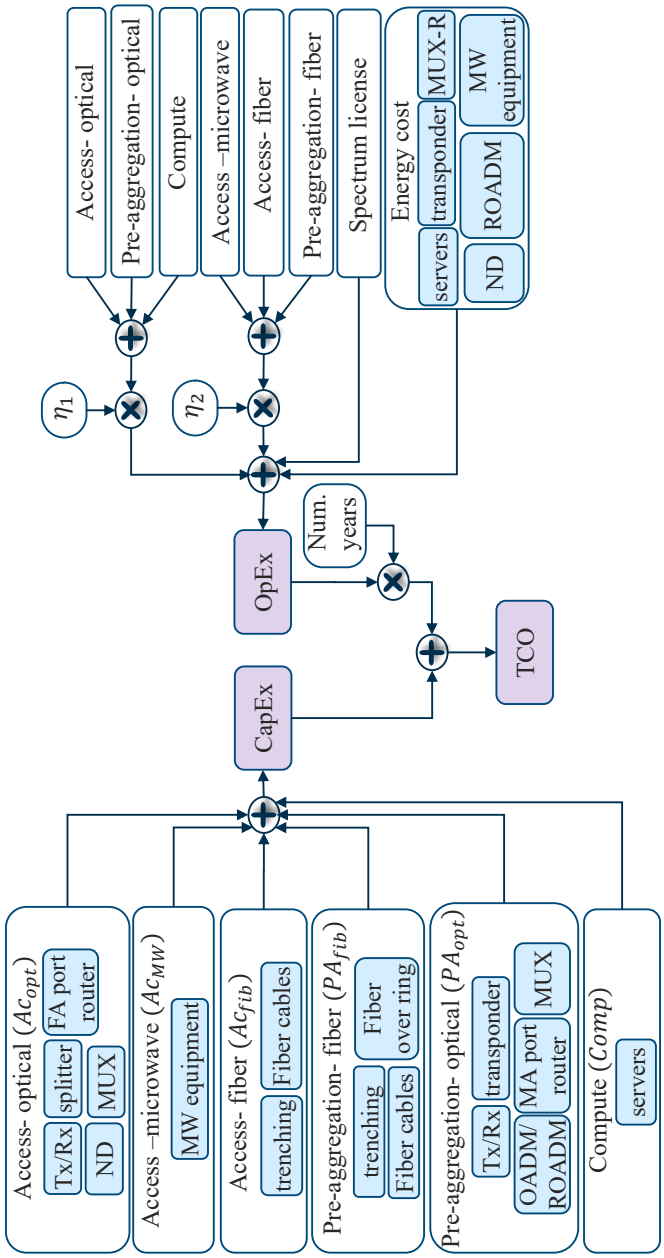


Figure 3.5: The proposed TCO model.

Connection availability characterization

We aim to evaluate the connection availability performance of the considered architectures. To achieve this, we calculate the connection availability between any cell site and its corresponding DC. Then, we make an average over connection availability values for all cell sites.

The generic availability modeling is explained in detail in Section 2.5. In this thesis, the connection between FA and MA is protected over the ring. Thus, its contribution to unavailability is negligible, and we ignore that segment in our calculations. Any of the components on the path from the cell site to the FA might fail. Thus, we need to consider their contributions to connection unavailability.

For example, consider the connection unavailability of $H^F_{NR,2}$. By using Eq. 2.2 and 2.3, the connection unavailability between a cell site and the DC can be estimated as the summation of unavailability of unprotected elements (Tx/Rxs, ND at the cell site, fiber cable, power splitter, multiplexer (MUX), ND at the FA, and OADM) plus the product of the unavailability of components that are protected (colored Tx/Rx plugged into ND to send traffic to OADM). The connection availability in the considered architectures can be then formulated as [41]:

$$A_{\text{conn}} = 1 - UA_{\text{conn}} = 1 - \left(\sum_{i \in \mathcal{NP}} UA_i + \sum_{(j,k) \in \mathcal{P}} UA_j \times UA_k \right), \tag{3.2}$$

UA_{conn} shows the connection unavailability from a given cell site to the related DC. UA_i is the unavailability of component i and can be calculated as in Eq. 2.1. A component belongs to the set of protected devices ($\mathcal{P}$) if at least one counterpart can take over as backup in case of failure (e.g., components connected in parallel in Fig. 2.7b belong to a protected set). Otherwise, the component belongs to the unprotected set ($\mathcal{NP}$).

To compare considered architectures, we define the average availability performance metric as follows:

$$P(A) = \frac{\text{Num. of Sites with } A_{\text{conn}} \geq A}{\text{Total Num. of Sites}} \times 100. \tag{3.3}$$

where A_{conn} is the connection availability between the cell site and the re-

spective DC. $P(A)$ measures the percentage of cell sites that can meet the connection availability requirement of A.

3.4 Performance Evaluation

We developed a custom Python-based framework that can mimic the deployment of the considered architectures. The framework is explained in detail in the following subsection. It gets different inputs (e.g., network dimension, service requirements, cost of components, type of urban area) and evaluates the TCO, latency, and availability performance metrics.

In **Paper A**, we considered an early-stage deployment of 5G, where the number of cell sites is not large enough to provide 5G coverage everywhere, and evaluated the TCO for an HLS option. The results presented in this section are based on **Paper B** and **Paper C**, which consider a mature 5G deployment stage with a sufficient number of cell sites.

We consider three geo-types (i.e., dense urban, urban, and sub-urban) with different numbers of cell sites, FA nodes, average link length, and required traffic capacity of MC/SC. These values are illustrated in Table 2 of **Paper B**, which are received from a system vendor and based on a real network deployment of a non-incumbent mobile operator in a city in South America. This thesis only shows the results for a dense urban geo-type. A summary of results for other geo-types is presented in Table 3.2, while the detailed discussions and results are available in **Paper B** and **Paper C**.

To evaluate TCO and availability performance, the cost, mean time to repair (MTTR), and mean time to failure (MTTF) values for optical components, microwave and mmWave band devices, NDs, and computing servers are shown in Table 3 in **Paper B**. The cost values are expressed in cost unit (CU), which is the cost of a 10 [Gb/s] gray Tx/Rx. We received this data from a vendor of microwave and optical equipment. The power consumption of active components is also shown in Table 3 in **Paper B** and is expressed in power consumption unit (PCU), corresponding to one transponder's power consumption.

The transmission rates of colored Tx/Rxs, gray Tx/Rxs, and transponder in Fig. 3.2 are assumed to be 25 [Gb/s], 100 [Gb/s], and 100 [Gb/s], respectively. We assume all Tx/Rxs in Fig. 3.3 have a rate of 10 [Gb/s]. In Figs. 3.4a and 3.4b, all Tx/Rxs have a rate of 25 [Gb/s] except the ones that are connecting

the NDs (located at the MC sites and feeder nodes) to the DC, which work at 100 [Gb/s].

We assume that fiber cables and microwave equipment are easier to maintain than other components and computing servers. Accordingly, in Fig. 3.5, $\eta_1 = 15\%$ and $\eta_2 = 5\%$. We calculate the TCO over five years of operation.

Performance evaluation framework

The framework is used for modeling the network and evaluating the TCO, latency, and availability performance of the considered architectures. The framework consists of different Python files (.py), shown in Fig. 3.6, along with their main functionality.

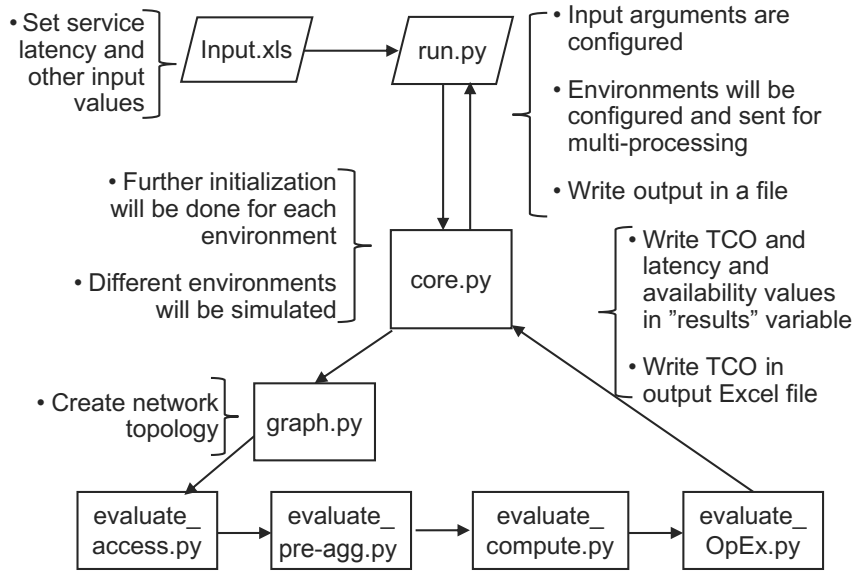


Figure 3.6: Framework workflow and the relation between different Python files.

Different inputs can be given to the framework (e.g., number of cells, network dimension, geo-type scenario, the intended architectures, cost of equipment, MTTF and MTTR of components, service latency requirements, etc.). Then, the framework creates different environments for faster processing (i.e., multi-processing), and each environment evaluates an architecture and a geo-

type. The framework has different modules to calculate the cost of the access segment, pre-aggregation segment, compute resources, and OpEx. The framework outputs are the TCO, latency, and availability performance for each architecture and geo-type. These values are written in an Excel file and a Python output file. The output files are then used to plot the required figures.

TCO analysis

Figure 3.7a shows the TCO breakdown for the HLS architectures (presented in Fig. 3.2 and 3.3). The microwave-based architecture (H_{NR}^{MW}) has lower TCO than the fiber-based ones ($H_{NR,1}^{F}$ and $H_{NR,2}^{F}$). The cost associated with the access segment in the microwave-based architecture ($Ac_{\mathrm{MW}} + Ac_{\mathrm{opt}}$) is lower than the cost of the access segment in the fiber-based ones ($Ac_{\mathrm{fib}} + Ac_{\mathrm{opt}}$). The cost of the pre-aggregation segment and compute servers are the same in the fiber- and microwave-based architectures as they are independent of chosen technology for the access. The TCO gain of H_{NR}^{MW} compared to $H_{NR,1}^{F}$ and $H_{NR,2}^{F}$ is higher in urban and sub-urban geo-types (shown in Table 3.2 and **Paper B**), as the average link length is longer in those areas resulting in higher fiber deployment cost.

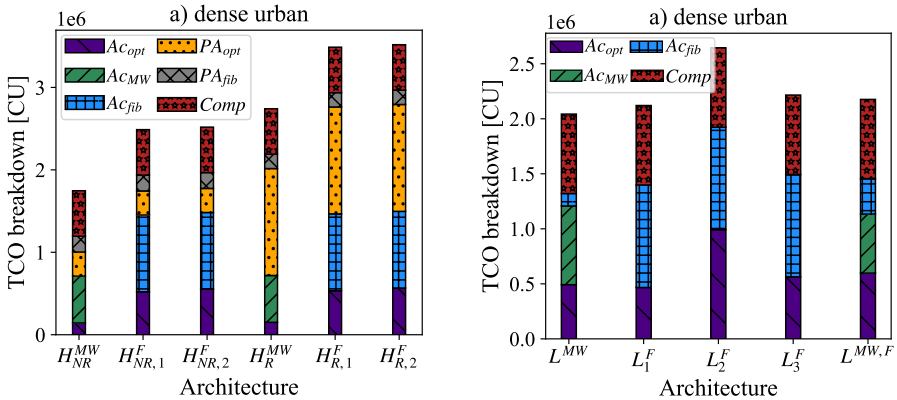
The TCO gains of reconfigurable (R) architectures (H_{R}^{MW} compared to $H_{R,1}^{F}$ and $H_{R,2}^{F}$) in Fig. 3.7a follow a similar trend as the non-reconfigurable (NR) ones. The absolute value of TCO in R architectures is higher because of more expensive components used in the pre-aggregation segment (i.e., ROADM, 100 [Gb/s] Tx/Rx, and 100 [Gb/s] transponders). However, the R architectures provide larger capacity and flexibility over the pre-aggregation segment.

The sensitivity analysis of results to the varying costs of microwave equipment and fiber trenching (due to the higher/lower labor cost and the ability to negotiate a reasonable price) is presented in **Paper B**.

The TCO breakdown of considered LLS architectures is shown in Fig. 3.7b. The TCO of microwave-based (L^{MW}) and hybrid fiber-microwave architectures ($L^{MW,F}$) is comparable to the TCO of the fiber-based architectures in dense urban geo-type. In dense urban areas, the average link length is typically short, which leads to a relatively low fiber deployment cost. This cost is balanced with the high price of required microwave devices with high capacities for LLS. The results for urban and sub-urban geo-types (presented in

Table 3.2, **Paper B**, and **Paper C**) show that L^{MW} has lower TCO compared to other architectures. The reason is the longer average link lengths in urban and sub-urban areas and, thus, the high cost of fiber deployment (Ac_{fib}). The benefits of $L^{MW,F}$ are more evident in urban and sub-urban geo-types (see Table 3.2). Indeed, $L^{MW,F}$ has lower costs compared to fiber-only-based architectures (by utilizing microwave technology whenever possible) and uses fiber wherever latency requirements cannot be met with microwave-based architecture (L^{MW}).

As for the HLS case, a sensitivity analysis of TCO is presented in **Paper B**, which investigates the dependence on negotiated price of microwave equipment and fiber trenching.



(a) TCO breakdown for HLS option.

(b) TCO breakdown for LLS option.

Figure 3.7: TCO evaluation: HLS option (non-reconfigurable (NR) and reconfigurable (R) architectures) and LLS option. Fig. 3.7a is reprinted with permission from **Paper B** ©Optica Publishing Group.

Latency evaluation

To evaluate the latency performance of the considered architectures, we look into two service categories, i.e., eMBB and URLLC (including URLLC-latency-sensitive (URLLC-S) and URLLC-latency-tolerant (URLLC-T)). The latency requirements (L_{HLS} and L_{LLS}) and the values of latency contributors are shown in Table 3.1. l_{BBU} and l_{OTA} are assumed to be different for eMBB

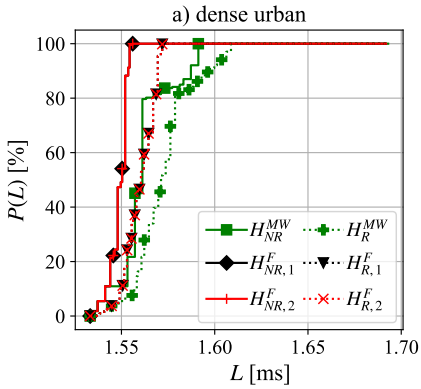
and URLLC services [52]–[54] to cope with their specific requirements. $l_{\text{prop}}^{\text{MW}}$ refers to latency due to the propagation on each microwave link. The propagation latency over fiber is calculated as the ratio between traversed distance and the speed of light propagation in fiber (v=2×10^8 [m/s]).

Table 3.1: Latency requirements and contributing values in [ms] for considered services [10], [43], [52], [53], [55].

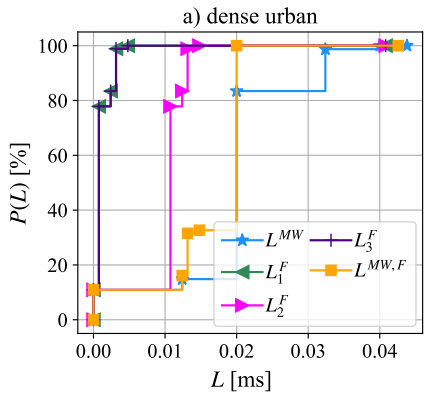
	latency requirements		latency values			
	L_{HLS}	L_{LLS}	l_{OTA}	l_{BBU}	$l_{\text{prop}}^{\text{MW}}$	l_{ND}
eMBB	4	0.1	0.5	1	0.02	0.01
URLLC-T	1	0.05	0.125	0.2	0.02	0.01
URLLC-S	0.5	0.025	0.125	0.2	0.02	0.01

Figure 3.8a shows the latency performance of HLS architectures when eMBB service needs to be provisioned. We observe that 100% of cell sites can meet the latency requirement of an eMBB service (i.e., 4 [ms]). Overall, the range of latency variation is small, leading to the conclusion that fiber- and microwave-based architectures have similar latency performance. The latency performance when provisioning the URLLC service is presented in Table 3.2 and **Paper B**. Results show that the considered architectures for HLS can meet the latency requirement of URLLC-T and URLLC-S (i.e., 1 [ms] [43] and 0.5 [ms]) except H_R^{MW}, where 11% of sites in the sub-urban area cannot meet the requirement of 0.5 [ms].

Figure 3.8b presents the latency performance of LLS architectures. The range of latency values is small, and all the considered architectures perform almost similarly. All architectures (except L^{MW}) can meet different latency requirements in Table 3.1 (i.e., L_{LLS} equal to 0.1 [ms], 0.05 [ms], and 0.025 [ms]). When using L^{MW}, a small percentage of sites (18%-37% depending on the geo-type) cannot meet the latency requirement of URLLC-S (0.025 [ms]). The hybrid fiber-microwave architecture ($L^{MW,F}$) is a suitable option that not only meets the 0.025 [ms] latency requirement for 100% of sites but also contains the TCO compared to fiber-only-based architectures. However, some operators may be willing to tolerate the slight decrease in latency performance in L^{MW}, in exchange for benefits such as faster installation, quicker time-to-market, and lower TCO [12], [14].



(a) Latency performance for HLS option with eMBB service.



(b) Latency performance for LLS option.

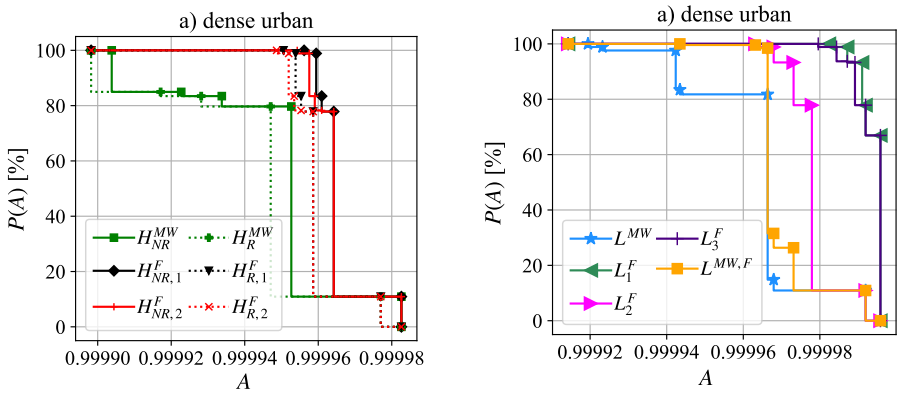
Figure 3.8: Latency performance evaluation for HLS and LLS options. Fig. 3.8a is reprinted with permission from **Paper B** ©Optica Publishing Group.

Availability evaluation

In the considered architectures, failure in fiber cable can occur because of fiber cuts and other physical damages during urban construction. The failure in the microwave can happen because of radio link failure and/or hardware failure. Radio link failures may occur when the received signal level falls below a certain threshold, which can be attributed to electromagnetic propagation conditions, including phenomena such as multipath and rain fading. To ensure high availability levels, ITU-R mandates that microwave links must be designed to meet a minimum availability of 0.99999 [56]–[58]. System vendors must comply with this standard when designing microwave links.

We evaluate the availability performance as explained in subsection 3.3. The MTTR and MTTF values of components are reported in **Paper B**. Figure 3.9a shows the availability performance of HLS architectures. According to 3GPP, the connection availability requirement of a TN should be in the [0.999, 0.9999999] range, depending on the requirements of services that need to be supported by the network [59]. Results show that the connection availability in the considered architectures for HLS is larger than 0.9999, and it is 0.99998 for 11% of cell sites. Moreover, the connection availability in the

LLS architectures (Fig. 3.9b) is higher than 0.9999, while a connection availability of 0.99999 can be provided by 11% - 93% of sites depending on the architecture option.



(a) Availability performance for HLS option.

(b) Availability performance for LLS option.

Figure 3.9: Availability performance evaluation for HLS and LLS options. Fig. 3.9a is reprinted with permission from **Paper B** ©Optica Publishing Group.

Other geo-types

As already explained throughout Section 3.4, Table 3.2 shows a summary of TCO and latency performance results for urban and sub-urban geo-types, which are explained in detail in **Paper B**. The microwave-based architectures have lower TCO than their fiber-based counterparts. In these areas, the link length is longer, on average, which results in higher fiber deployment costs.

In terms of latency performance, L^{MW} can meet the latency requirement of URLLC-S (i.e., 0.025 [ms]) in 71% and 63% of cell sites in urban and sub-urban geo-types, respectively. On the other hand, the hybrid fiber-microwave architecture ($L^{MW,F}$) can meet the latency requirement in 100% of sites (by leveraging fiber where needed). At the same time, its TCO is lower than the fiber-only-based architectures.

It should be pointed out that the results discussed so far are for a non-incumbent operator. We also run the simulations for an incumbent operator.

Table 3.2: TCO of considered architectures for urban and sub-urban geo-types (values are in cost unit [CU] divided by 10^6). Percentage of sites in urban and sub-urban that meet latency requirements of eMBB, URLLC-T, and URLLC-S. Ur: urban, SUr: sub-urban.

Architecture	TCO		eMBB		URLLC-T		URLLC-S	
Geo-type	Ur	SUr	Ur	SUr	Ur	SUr	Ur	SUr
H_{NR}^{MW}	1.7	0.7	100%	100%	100%	100%	100%	100%
$H_{NR,1}^{F}$	2.9	1.5	100%	100%	100%	100%	100%	100%
$H_{NR,2}^{F}$	2.9	1.5	100%	100%	100%	100%	100%	100%
H_{R}^{MW}	3	1.3	100%	100%	100%	100%	100%	89%
$H_{R,1}^{F}$	4.2	2.1	100%	100%	100%	100%	100%	100%
$H_{R,2}^{F}$	4.2	2.1	100%	100%	100%	100%	100%	100%
L^{MW}	1.9	0.8	100%	100%	100%	100%	71%	63%
L_{1}^{F}	2.5	1.3	100%	100%	100%	100%	100%	100%
L_{2}^{F}	2.8	1.4	100%	100%	100%	100%	100%	100%
L_{3}^{F}	2.6	1.4	100%	100%	100%	100%	100%	100%
$L^{MW,F}$	2.3	1.2	100%	100%	100%	100%	100%	100%

The number of cell sites, FA nodes, and MAs differ from where the operator is non-incumbent. Moreover, the incumbent operator owns fiber over pre-aggregation rings but needs to deploy fibers to connect cell sites to the FA. The simulation results (TCO, latency, and connection availability) for the incumbent scenario have a similar trend as in the non-incumbent case. The only difference is the absolute value of TCO, latency, and availability performance metrics. Part of the results for an incumbent operator is shown in **Paper C**.

3.5 Summary

This chapter presents a holistic framework to assess the performance of several microwave- and fiber-based TN architectures using HLS and LLS options. The architectures are compared regarding TCO, latency, and connection availability in three geo-types (dense urban, urban, and sub-urban). Three service types (eMBB, URLLC-T, and URLLC-S) with different requirements are con-

sidered.

The results in the HLS option show that the microwave-based architecture has lower TCO than the fiber-based architecture. The TCO gains are the largest in sub-urban geo-type than other urban areas. Moreover, TCO gains of urban are larger than the dense urban area. The reason is that the average link length increases going from a dense urban to a sub-urban area, resulting in higher fiber deployment costs. All the considered architectures have almost similar latency performance and can meet the requirements of eMBB and URLLC services. The only exception is our considered reconfigurable-microwave-based architecture, which cannot meet the latency requirement of URLLC-S in 11% of sites in the sub-urban area. The investigated architectures also have similar connection availability performance within the range specified by 3GPP.

In the LLS option, microwave-based architecture has comparable TCO to fiber-based ones in dense urban. The reason is that the short average link length in dense urban results in the almost similar cost of fiber deployment and microwave equipment. On the other hand, in urban and sub-urban areas, microwave shows consistent TCO gain as the fiber deployment costs are very high due to longer average link length than in dense urban. All the considered architectures can meet the latency requirement of eMBB, URLLC-T, and URLLC-S services, except microwave-based architecture that cannot satisfy the requirement of URLLC-S in 18%-37% of sites (depending on the geo-type). The strict latency requirements of URLLC-S can be met in 100% of sites using a hybrid fiber-microwave architecture. Indeed, a hybrid fiber-microwave architecture eliminates the need for multi-hop microwave links. Instead, to leverage microwave cost benefits, this technology is used in the hybrid architecture only for the sites where possible.

In conclusion, determining the optimal solution for TN deployment is a complex task. Nonetheless, the findings of our study indicate that microwave technology is a viable solution for 5G and beyond transport networks, as it can satisfy diverse 5G service requirements. As a result, when fiber deployment is not feasible within operators' cost and time objectives, microwave technology can be considered an effective alternative.

CHAPTER 4

Resilient Network Design Using Shared Protection Resources

As mentioned in Chapter 1, operators are looking for strategies to guarantee the resiliency of their network against failures. In this chapter, we present a strategy that ensures the survivability of any service running over a midhaul network in a single failure scenario while maximizing resource efficiency. The proposed strategy selects primary and backup resources and tries to maximize sharing of the backup resources as much as possible. The network resiliency against failures in the midhaul nodes/links and DC nodes is ensured.

First, we present a summary of existing works in the literature. Next, we define the considered system architecture. Then, we present the use case definition and explain the proposed strategy. Finally, the performance evaluation and results are illustrated.

4.1 Literature Review

Designing a resilient C-RAN architecture has been addressed extensively in the existing literature. The work in [60] proposed an algorithm for baseband

unit (BBU) pool placement in C-RAN so that survivability against single BBU pool failure is guaranteed. The presented algorithm works in a two-step fashion. First, it deploys the minimum number of primary and backup BBU pools and selects the connectivity path between remote radio units (RRUs) and their primary and backup BBU pools. Then, it tries to maximize the sharing of the backup BBU ports among RRUs to reduce the total cost. The authors in [61] proposed three methods for joint BBU pool placement and traffic routing problems. The three approaches presented in [61] include dedicated path protection, dedicated BBU protection, and dedicated BBU and path protection. They introduced integer linear programming formulations to minimize the number of BBU pools, the number of used wavelengths, and the amount of compute resources for baseband processing. The work in [62] formulates an integer linear programming problem where an operator aims to choose the BBUs from different cloud providers. The objective is a weighted sum of three optimization goals, including minimizing the consumed processing power at the BBU pool, maximizing resiliency (in which they define resiliency as a function of failure probability of BBU pools), as well as the amount of traffic load that can be handled from RRUs. The work in [63] presented a survivable 5G network architecture against single link failure using dedicated path protection. The network design is modeled as an integer linear programming problem to minimize deployment costs. The authors in [64] proposed a scheme to reduce the high bandwidth requirements for protection in C-RAN architecture. Their results show that the total required capacity in the optical network can be reduced by 33% using their proposed strategy.

Most of the works in the literature on resilient C-RAN design (e.g., [61], [63], [64]) use dedicated protection, which is not a resource-efficient approach compared to shared protection. As Section 2.4 explains, shared protection requires fewer backup resources than dedicated protection. Some works (e.g., [60]) considered sharing but only among backup BBU ports, while sharing backup compute and connectivity resources can further improve the cost-efficiency in network design. Moreover, these studies operate in a two-step fashion to benefit from sharing. First, the location of primary and backup BBU pools are selected without considering the impact of this choice on the potential of sharing backup resources. Then, in the second stage, the backup resources are shared as much as possible. On the other hand, the cost savings can be further increased if the shareability potentials are considered during

BBU pool placement.

In the next section, we present a network architecture to evaluate the cost savings of a strategy that considers the shareability potential when selecting BBU pool locations.

4.2 Network Architecture

Operators can choose among different RAN architectures depending on their needs, as explained in Section 2.3. We consider an architecture with fronthaul, midhaul, and backhaul segments, i.e., referred to as a hybrid cloud radio access network (H-CRAN) [65] shown in Fig. 4.1. H-CRAN is composed of three elements: remote radio unit (RRU), radio aggregation unit (RAU), and radio cloud center (RCC) [66]. The segments between RRU-RAU, RAU-RCC, and RCC-core are referred to as fronthaul, midhaul, and backhaul, respectively. Several RRUs are connected to an RAU node, while RAU nodes are connected to RCC nodes. We assume a functional split option 2 between the RAU and RCC and a functional split option 7.2x between the RRU and RAU. Accordingly, as explained in Section 2.3, among 5G protocol stack functions, RRC and PDCP are deployed in RCC nodes, while RLC, MAC, and high-PHY are implemented in RAUs. Low-PHY and RF are processed in RRUs [6], [15]. Each RAU i requires several server units, i.e., s_i, in an RCC node. These servers are needed for processing baseband functions and services requested by RRUs connected to a given RAU. Therefore, s_i depends on the number of connected RRUs to RAU i.

We explain the use case and the proposed resilient design strategy in the next section.

4.3 Use Case Definition

A failure can occur in various nodes/links of the H-CRAN architecture with different impacts on the number of affected users. If a failure happens in an RCC node, many users will be impacted as each RCC node covers a large geographical area. Similarly, if a node/link in the midhaul segment fails, many RAU nodes and users will experience service disruption. Accordingly, for designing a resilient H-CRAN, considering resiliency in the presence of failures in midhaul nodes/links and RCC are critical [15], [66]. For this reason,

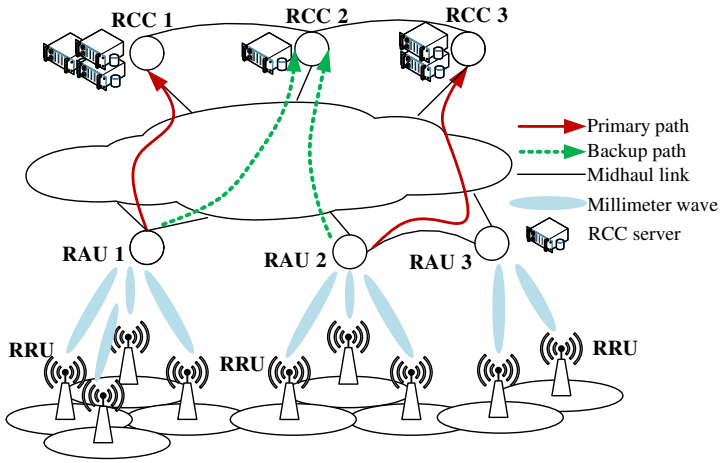


Figure 4.1: An example of an H-CRAN architecture. RAU nodes are connected to RCC nodes through midhaul segments. RRUs are connected to the RAU nodes via the fronthaul segment. **Paper D** ©2019 IEEE.

we consider them in our study.

We assume a single failure scenario, i.e., the failure in the node or link is repaired before another failure occurs in the network. The objective is to guarantee the survivability of services running on the network against a single failure. We achieve this target by assigning one primary and one backup RCC node to each RAU node, connected via two node-disjoint paths in the midhaul, referred to as primary and backup paths. Assigning the RCC nodes and connectivity paths should be done such that 1) the service latency requirement is met, and 2) resource usage, and consequently, deployment costs are minimized. We assume that the number of hops over the primary and backup connectivity paths is constrained to a maximum allowable value to meet the latency requirement. We define cost as the sum of the deployment cost of RCC nodes, server units within RCC, and connectivity units as follows:

$$Cost = N_{RCC} \cdot C_{RCC} + N_{Ser} \cdot C_{Ser} + N_{Conn} \cdot C_{Conn}, \tag{4.1}$$

where N_{RCC} is the number of deployed RCC nodes, C_{RCC} is the cost of deploying one RCC node (expenses associated with construction work, cooling,

power modules, and other equipment in RCC), N_{Ser} is the total number of required server units, C_{Ser} is the cost of one server unit, N_{Conn} is the total number of connectivity units in the midhaul segment, and C_{Conn} is the cost of one connectivity unit.

Minimizing the deployment cost can be obtained by maximizing sharing of backup resources (i.e., server units in the backup RCC nodes and connectivity units over the backup path) as explained in Section 2.4. If two conditions are met, two (or more) RAU nodes can share backup server units and/or connectivity units on the backup path. First, their primary server units must be deployed in different RCC nodes. Second, their primary connectivity paths must have no common links/nodes (i.e., node-disjoint). We refer to these two prerequisites as *sharing conditions*.

To minimize cost when choosing primary and backup RCC nodes and the connectivity paths, we proposed a heuristic algorithm called shared-path shared-compute planning (SPSCP). The intuition behind this algorithm is to check, for each RAU node, all the options for primary and backup RCC nodes and choose the one with the lowest cost according to the cost function defined in Eq. 4.1. The details of this algorithm are explained in Fig. 4.2.

In SPSCP, the set of transport network nodes can be categorized into two parts, i.e., the set of nodes where RAUs are located (referred to as set R_a) and the remaining nodes, where RCCs can be deployed (referred to as set R_c). First, we sort all nodes in R_a based on the increasing order of their nodal degree to get a set referred to as $\mathcal{A}_s$. Then, for a given number of hops (h), we calculate the *combined degree* of nodes in set R_c. We define the *combined degree* of a given node as the summation of its nodal degree and the number of RAU nodes within h hops from the node. For each RAU in $\mathcal{A}_s$, we find $\mathcal{P}'$, which is a subset of nodes in R_c located within h hops from the RAU and sorted based on the decreasing order of their combined degree. Then, we set the minimum value of cost to infinity ($C_{min} = \infty$). For all possible primary and backup options for the RCC node (i.e., k, m, such that $k \in \mathcal{P}'$ and $m \in (\mathcal{P}' - k)$), we calculate deployment cost ($C_{m,k}$) using Eq. 4.1. We select the option (k, m) with minimum cost as the ultimate choice of primary and backup RCC, respectively. If the backup server units and/or backup connectivity units are shared, their cost is only counted once in $C_{m,k}$. The backup server and connectivity units can be shared if the *sharing conditions*, which were already explained, are met. We use the shortest path algorithm to

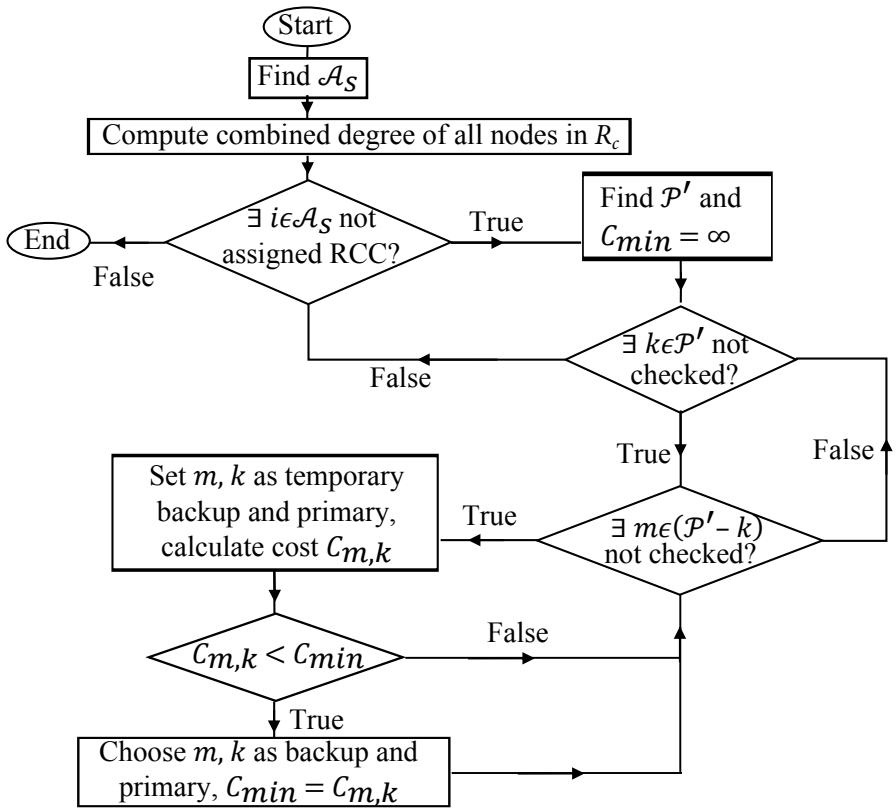


Figure 4.2: The flowchart of SPSCP strategy [**Paper D**] [18].

find primary and backup connectivity paths. This process is repeated until a primary and a backup RCC (and the related connectivity paths) are assigned to all RAU nodes in $\mathcal{A}_s$. The performance of this algorithm is evaluated in the next section.

4.4 Performance Evaluation

In this section, we illustrate the performance of the SPSCP strategy. We consider a transport network with 38 nodes shown in Fig. 4.3 where 40% of

nodes are chosen randomly with a uniform distribution to host RAU nodes (set R_a). We assume that the number of required server units by an RAU (s_i) is chosen uniformly within the range $[5, 15]$. We further assume that the deployment cost of an RCC (C_{RCC}) is 100 cost units [CU]. C_{Ser} and C_{Conn} are assumed to be much smaller than C_{RCC} to reflect the fact that the costs of one server unit and one connectivity unit are lower than RCC deployment. To investigate the impact of the relative cost of server and connectivity units, we evaluated two cases, where in the first case, $C_{Conn} = 1, C_{Ser} = 5$, and in the second one, $C_{Conn} = 5, C_{Ser} = 1$. Indeed, C_{RCC}, C_{Ser}, and C_{Conn} are sort of weighting parameters in the objective function (Eq. 4.1). The maximum number of allowable hops (h) is varied between 2 and 10.

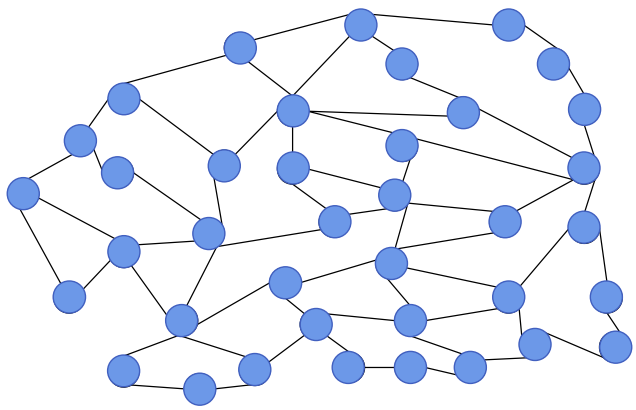


Figure 4.3: The midhaul network with a mesh topology and 38 transport network nodes. **Paper D** ©2019 IEEE.

The performance of SPSCP is evaluated against three benchmark algorithms. The first one is referred to as resource duplication (RD), where each RAU is connected to a primary and a backup RCC node via two node-disjoint connectivity paths. The pairing is done without the possibility of sharing any backup resources [60]. The second benchmark is referred to as preliminary resource sharing (PRS). PRS is based on the initial RD solution, on top of which, we share backup resources (where possible) without modifying the initial pairing between RAUs and their backup RCC nodes [60]. The third benchmark is referred to as reconfiguration and improved resource sharing (RIRS). RIRS revisits/changes the initial pairing decided by RD between RAUs and

their backup RCC nodes (and related connectivity paths) in order to maximize sharing.

It should be pointed out that PRS and RIRS try to share backup resources in a second stage after the RCC nodes are deployed and the connectivity paths are specified. Therefore, they do not consider the potential of sharing backup resources when making the initial pairing between RAU and RCC nodes. On the other hand, SPSCP considers the potential of sharing (and its cost benefits) when deploying RCC nodes.

Figure 4.4 shows the cost savings of SPSCP and two benchmark methods (i.e., PRS and RIRS) compared to RD as a function of the number of allowable hops (h). The connectivity and server unit costs are assumed to be $C_{Conn} = 1, C_{Ser} = 5$ cost units ([CU]), respectively. SPSCP can achieve 16% to 28% cost savings with respect to RD for different values of h. It can also achieve 28% and 23% cost savings compared to PRS and RIRS, respectively, for $h \geq 6$.

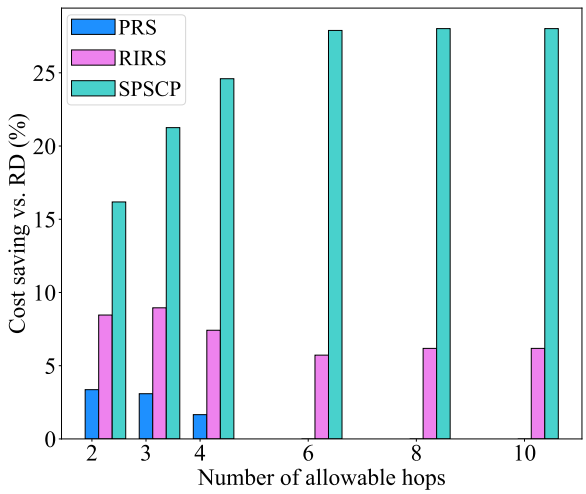


Figure 4.4: Total cost savings for PRS, RIRS, and SPSCP compared to RD when $C_{RCC} = 100, C_{Conn} = 1, C_{Ser} = 5$. **Paper D** ©2019 IEEE.

It is interesting to note in Fig. 4.4 that the cost savings of SPSCP increase by relaxing the hop count constraint. The reason is that SPSCP deploys primary and backup RCC nodes on a few transport network nodes (i.e., concentrating them to reduce the deployment cost of RCC). At the same time, it considers

the sharing potential and its cost benefits. On the contrary, the cost savings of RIRS compared to RD decreases by relaxing the hop count constraint. Indeed, RIRS does not consider the potential shareability of resources in the initial deployment of RCC nodes. Thus, to reduce RCC cost, RIRS tries to concentrate RCCs on a few transport network nodes as much as possible. This, in turn, reduces the possibility of sharing backup servers and makes RIRS more comparable to RD.

The results in Fig. 4.4 can be confirmed by looking at the breakdown of cost in Fig. 4.5 in terms of the cost of primary/backup servers, primary/backup connectivity, and RCC for the four considered methods.

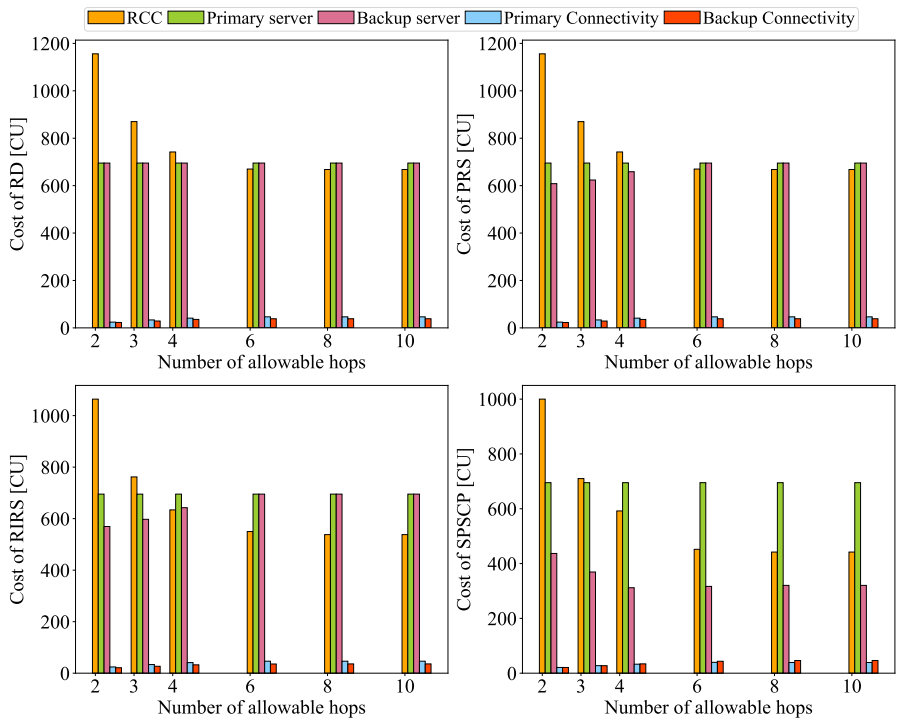


Figure 4.5: Breakdown of cost when $C_{RCC} = 100, C_{Conn} = 1, C_{Ser} = 5$. **Paper D** ©2019 IEEE.

As shown in Fig. 4.5, using RIRS, the primary and backup servers' cost is equal when $h \geq 6$. It means that backup servers cannot be shared (similar to

what happens in the RD approach), resulting in lower cost savings compared to RD. Moreover, sharing the backup connectivity resources has a negligible impact on the overall cost. On the other hand, when $h \geq 6$, the cost of required backup servers in SPSCP is much lower than the one in RD, which is obtained thanks to the possibility of sharing these resources. Additionally, as illustrated in Fig. 4.5, the main cost savings of SPSCP are the result of shared backup servers and a large reduction in the number of required RCC nodes. It is worth noting that the total cost of all methods decreases as the hop count constraint is relaxed.

Figure 4.6 presents the cost savings of PRS, RIRS, and SPSCP compared to RD when connectivity and server unit costs are assumed to be $C_{Conn} = 5, C_{Ser} = 1$ [CU], respectively. This figure shows SPSCP has similar trends as presented in Fig 4.4 with respect to RD, i.e., the cost saving increases by relaxing hop count constraint and can reach up to 24%.

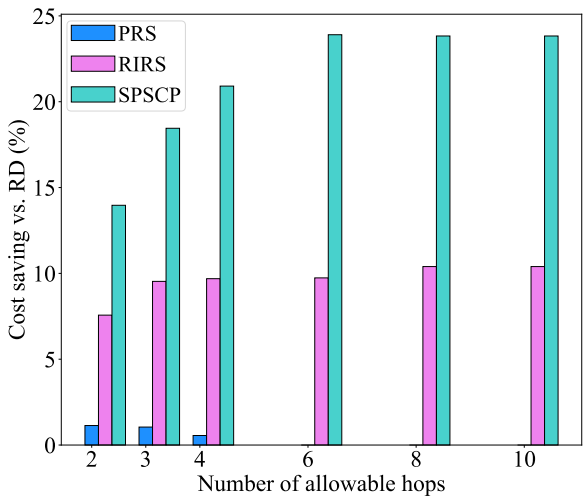


Figure 4.6: Total cost savings for PRS, RIRS, and SPSCP compared to RD when $C_{RCC} = 100, C_{Conn} = 5, C_{Ser} = 1$. **Paper D** ©2019 IEEE.

Figure 4.6 illustrates how the cost savings of RIRS compared to RD increase as the hop count constraint is relaxed. This can be explained by looking at Fig. 4.7 (i.e., breakdown of cost when $C_{Conn} = 5, C_{Ser} = 1$). In this case, C_{Ser} is smaller compared to that shown in Fig. 4.5, and the server cost is not

the primary driver of the total cost in Fig. 4.7. Consequently, the inability to share backup servers (at hop count $h \geq 6$) has a negligible impact on the cost savings of RIRS. On the other hand, SPSCP can still share backup servers among RAUs, while other benchmark methods cannot effectively benefit from sharing backup servers. Moreover, another critical driver of cost savings in SPSCP is the substantial reduction in the cost associated with RCC nodes (for the same reason as in Fig. 4.5).

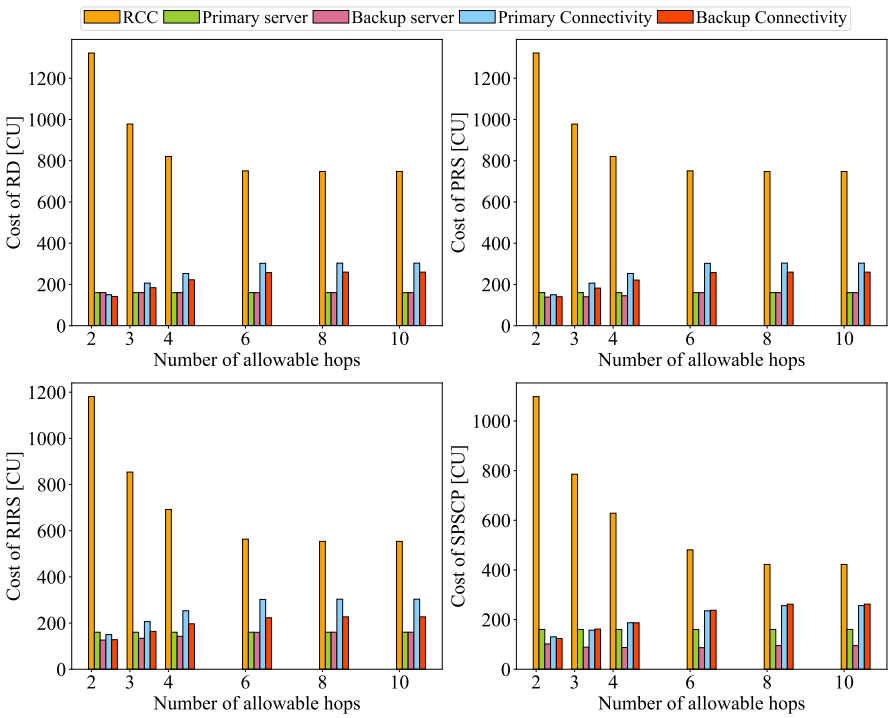


Figure 4.7: Breakdown of cost when $C_{RCC} = 100, C_{Conn} = 5, C_{Ser} = 1$. **Paper D** ©2019 IEEE.

4.5 Summary

This chapter presents a strategy for designing a resilient midhaul network in an H-CRAN architecture with a minimal deployment cost. The proposed

strategy, referred to as shared-path shared-compute planning (SPSCP), can guarantee the survivability of any service when a single failure occurs. The failures in the RCC nodes and nodes/links of the midhaul segment are considered.

SPSCP assigns a primary and a backup RCC node to each RAU node, connected via two node-disjoint paths, while trying to minimize the cost. To achieve this, SPSCP considers the shareability potential of the backup servers and connectivity resources when deploying RCC nodes and tries to maximize sharing.

Evaluation results show the performance of the proposed strategy compared to three benchmark methods. SPSCP can obtain 28% cost savings compared to a conventional approach that assigns dedicated backup resources. Moreover, SPSCP can yield up to 23% cost savings compared to another approach that tries to share resources only in the second stage after RCC nodes are deployed.

CHAPTER 5

Resource-efficient Service Deployment Using Compute Off-Loading

As mentioned in Chapter 1, eMBB, URLLC, and mMTC are three service categories in 5G. We expect new services such as 6 degrees of freedom (6-DoF) virtual reality (VR), holographic communications, and tactile internet in beyond 5G networks. With the emergence of new services, the latency, availability, compute, and connectivity resource requirements are getting more demanding [1], [4], [67]. Operators seek service provisioning approaches to use their infrastructure resources efficiently and maximize their profit while meeting service requirements.

This chapter evaluates the benefits of deploying services in central DCs (where possible) to improve resource efficiency and profit in a dynamic service provisioning scenario. First, we briefly overview existing works on the dynamic provisioning of 5G services in the literature. Next, we explain the intuition behind a centralized service deployment and discuss strategies for mitigating its limitations. Then, we illustrate the considered architecture and present the proposed strategy. We define the profit model and briefly explain the contributions to latency and availability calculation in this study. Finally, we show the simulation results and summarize the chapter.

5.1 Literature Review

Network infrastructures have a limited amount of connectivity and compute resources. These limited resources should be used efficiently to successfully provide services for many users while stringent latency and availability demands are met. Operators can take advantage of the varying requirements of different services to exercise flexibility and deploy them in different network locations. Moreover, operators are looking for strategies to maximize the profit achieved during network operation. In other words, they should accept as many service requests as possible (to generate higher revenue) with minimum resource usage (to have lower costs). Finding resource-efficient 5G service provisioning methods has gained interest from academia and industry.

The work in [36] presented a cloud-based service provisioning approach while the network resiliency against failures is guaranteed. Clients request IT resources (mostly storage and compute resources) in a cloud-based service. The proposed survivable strategy is based on restoration and benefits from cloud service relocation and differentiation. The service relocation is used to improve restoration performance, while the service differentiation is used to give proper attention to critical services when assigning backup resources. The objective of integer linear programming formulation and the heuristic approach proposed in [36] is to minimize the number of relocated cloud services and the average service downtime. The authors in [68] proposed a service provisioning solution to minimize the usage of fiber, processing, and storage resources. They considered several 5G and beyond services with given latency, compute, and storage requirements. The deployment of services was across hierarchical DCs (i.e., edge DCs, metropolitan DCs, and core DCs). The work in [69] developed a service provisioning algorithm considering a wireless-optical converged network architecture. The algorithm's objective was to optimize the usage of optical and wireless resources while fulfilling the services' specific delay and bandwidth requirements. The authors in [70] proposed a strategy to reduce the service provisioning time. Indeed, the service provisioning time (i.e., the time it takes to respond to a service request) must be low in a dynamic scenario. The presented strategy in [70] delays releasing the optical channel when the channel no longer carries any traffic so that it can be used for upcoming service requests.

Although the works on dynamic provisioning of 5G services in the literature (e.g., [36], [68]–[70]) have addressed many critical challenges, there is a gap in

research to maximize resource efficiency while considering both latency and availability constraints. The mentioned works above on service deployment need to evaluate the profitability of their proposed approach. In fact, operators are interested in enhancing their profit while provisioning services.

Regarding profit analysis, the authors in [17] discussed that operators must incur significant expenses to upgrade their network while the revenue per user is reducing. The high cost negatively impacts operators' profitability and ability to adapt to new standards. The authors investigated the effect of using SDN and network virtualization. They showed that leveraging these technologies (on top of classical architecture) can result in significant cost savings and increased profit. The work in [71] studied maximizing the profit of both service and server providers by introducing a distributed service deployment algorithm. The users request services with high QoS and large capacity requirements from the service provider, which, in turn, rents bandwidth and storage resources from the server provider. The presented approach incorporates bandwidth and caching costs simultaneously for the service deployment. The results indicate that the joint optimization approach yields favorable outcomes for overall profit. In [72], a framework to deploy 5G services was proposed to efficiently utilize compute and memory resources. The presented framework monitors service requirements, compute resource usage, and memory resource utilization to decide on the service placement. It results in higher revenue than the benchmark (i.e., random service deployment). The authors in [73] introduced a framework for deploying content delivery services to maximize profit. Their framework considers connectivity and compute resource utilization to enhance profit while meeting the service latency and bandwidth requirements. The framework considers service reconfiguration to improve profit. It also evaluates the impact of the penalty on overall profit due to violating latency requirements during the reconfiguration process. The work in [74] presented a service admission policy (where traffic prediction is used to accept service requests) and evaluated the profit of an infrastructure provider. The results indicate that, by using traffic prediction and considering future service requirements, the penalty due to service degradation can be mitigated, and profit is increased. The works in [71]–[74] evaluated profit focusing on capacity, compute, or latency constraints without considering service availability requirements. On the contrary, future services have stringent availability requirements [3], [4], which must be addressed in profitability assessment.

5.2 The Compute Off-loading Concept

Utilizing compute resources in the network can be improved by centralizing service processing. To evaluate this intuition and possible cost benefits of centralized service processing, in **Paper E**, we considered a simple TN architecture as in [44]. The TN architecture is shown in Fig. 5.1, composed of local, provincial, regional, and national segments. The DCs can be deployed in any of these segments. The users are connected to access points (APs), which, in turn, must be connected to a DC, where the application server (AS) for service processing is deployed. The traffic from the APs is aggregated at the local aggregation point and sent to the provincial segment. Afterward, the traffic is aggregated at the intersection of the provincial and regional segments and sent over the regional segment (and possibly through the national segment) until it reaches the DC location.

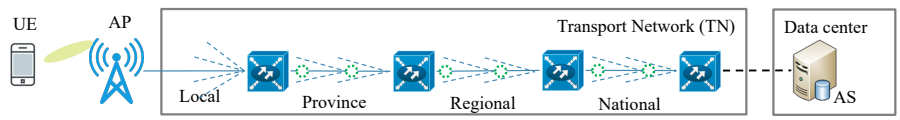


Figure 5.1: A network composed of 4 tiers (local, provincial, regional, and national). Reprinted with permission from **Paper E** ©OSA 2019.

Deploying DCs close to AP (e.g., in the local segment) translates into traversing a shorter connectivity path, but each DC can handle a limited number of APs. On the other hand, if DCs are placed far from APs (e.g., in the regional segment), they can provide service for a larger number of APs. Compute off-loading to the large and central DCs has various other advantages such as: 1) the resources in the edge DCs can be saved for services with stringent latency requirements [71], 2) central DCs usually have abundant compute resources, 3) centralized service processing has a lower cost because of economy-of-scale [22], [23], 4) higher tier TN segments can be used which allows multiplexing of traffic into fewer channels [24], and 5) compute resource utilization can be improved as the compute resources can be shared among many users/operators/services, which reduces energy consumption as well [22]. However, longer distances should be traversed to reach large DCs in higher-tier transport network segments. The service latency and availability requirement restrict the distance that can be traversed to reach these cen-

tral DCs. Nothing can address strict latency constraints other than deploying DCs close to the AP. However, as explained in Section 2.4, the availability performance can be improved by adding a backup path that can be used in case of failure on the primary path. The extra connectivity resources used on the backup path might reduce the cost benefits of centralized deployment. Therefore, it is essential to evaluate the cost benefits of centralized service deployment when adding a backup path.

As already explained, we assume a simple scenario for our initial assessment. We consider four 5G use cases with various latency and availability requirements discussed in **Paper E**. The use cases include 1: collaborative gaming, 2: remote control for smart manufacturing, 3: discrete automation, and 4: process automation/monitoring. Different requirements of these use cases result in different allowable distances between the AP and the DC for each use case. Accordingly, each use case can be deployed in a DC in a specific segment of the TN, leading to different costs for the considered use case. The cost is a function of the price of TN equipment and the DC.

By adding a backup path, the maximum allowable distance between the AP and the DC can be increased (as shown in Fig. 2 in **Paper E**). Therefore, the DC can be located in a higher-tier TN segment.

Figure 5.2 shows the cost savings of a centralized deployment (by adding a backup path) for the considered use cases as a function of γ. In this figure, γ is defined as a parameter indicating the cost efficiency of a national DC compared to regional, provincial, and local DCs. Indeed, the higher the γ, the less cost-efficient the regional, provincial, and local DCs are. Our initial evaluation confirms that adding a backup path allows for deploying DCs in the centralized locations, which results in cost savings. As use cases have various latency and availability requirements, the impact of adding a backup path on maximum allowable distance and, accordingly, on cost saving is different among use cases.

By increasing γ in Fig. 5.2, we observe that the cost savings of using large DCs are also increasing. For example, in $\gamma = 6$, the regional, provincial, and local DCs have the least cost-efficiency than other values of γ in the figure. Therefore, in $\gamma = 6$, using national DCs (which becomes possible by adding a backup path) leads to the highest cost savings (up to 74%).

Analyzing the trade-off and cost benefits of centralized deployment is crucial for network operators in dynamic service provisioning beyond 5G scenarios.

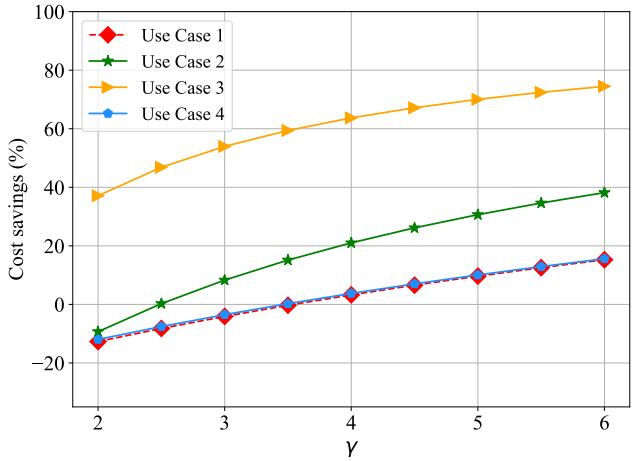


Figure 5.2: Cost savings of centralized service deployment vs. γ for four 5G use cases [**Paper E**].

In particular, we propose an approach for dynamic service provisioning, referred to as resource-efficient provisioning (REP). REP is a backup-enhanced compute off-loading strategy, i.e., it adds a backup path to encourage centralized service processing. In the following sections, we explain REP and assess operators' profit.

5.3 Dynamic Service Provisioning with Backup-Enhanced Compute Off-loading

We define *service provisioning* as the process of selecting a DC with enough compute resources to deploy the AS (for running the service). This process also involves choosing a connectivity path with enough capacity between the AP and the assigned DC. To ensure the service latency and availability requirements are met while optimizing resource efficiency, it is essential to carefully choose the DC and connectivity path.

In this section, we present the network architecture considered for service provisioning, and then, we introduce REP as the backup-enhanced compute off-loading method.

Network architecture

We consider the network architecture presented in Fig. 5.3, which uses wavelength division multiplexing (WDM) technology [75].

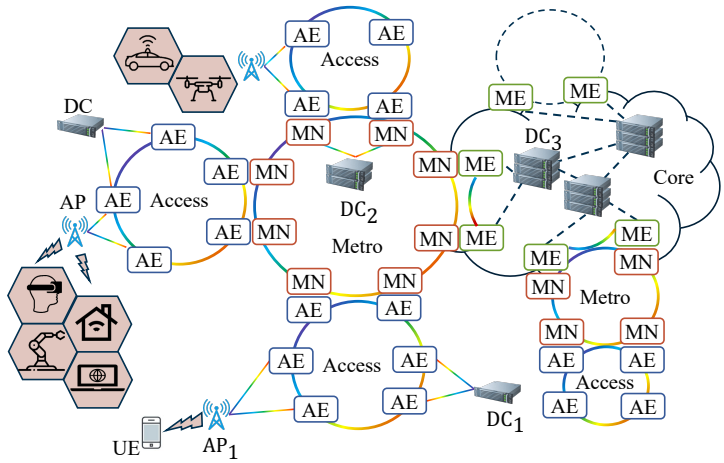


Figure 5.3: Network architecture with three tiers, i.e., access, metro, and core. The user equipment (UE) is connected to an access point (AP). Data centers (DCs) with different capacities are deployed throughout the infrastructure. AE: access edge, MN: metro node, ME: metro-core edge. Reprinted from **Paper G**.

The network is composed of three tiers, i.e., *access*, *metro*, and *core*. The access rings (composed of access edge (AE) nodes) are connected to metro rings (through metro nodes (MNs)), which, in turn, are connected to the core segment (via metro-core edge (ME) nodes). This type of network architecture aligns with the reference network presented in [17]. Networking devices perform the traffic aggregation and grooming in MN and ME nodes. We assume a fixed number of wavelengths in the rings. The access, metro, and core rings work at different transmission rates. DCs in various network segments have compute resource capacities specific to that segment. Core DCs are equipped with more compute resources than metro DCs, which, in turn, have a larger amount of compute resources than DCs in the access segment.

End users (e.g., UE, robots, machines, sensors, etc.) are connected to APs in the access segment. These users request different types of services (e.g.,

media streaming (MS), 6 degrees of freedom (6-DoF) virtual reality (VR), massive Internet-of-things (mIoT), vehicle-to-X (V2X), smart factories, etc.) with different latency, availability, compute, and connectivity requirements [3], [76]. A DC with enough compute resources must be selected to handle user service requests. The user must be connected to the selected DC by reserving sufficient connectivity resources. The selection of the DC and the connectivity path must be performed such that the service latency and availability requirements are met, and the resource efficiency is maximized. In the following section, we explain such a service provisioning approach.

Resource Efficient Provisioning (REP)

This section describes the resource-efficient provisioning (REP) strategy proposed in **Paper F**. All the variables and parameters used in this chapter are defined in Table 5.1.

Let us assume service request j is originating in AP δ, with latency, availability, compute, and data rate requirements defined by (L_j, A_j, S_j, R_j). REP selects a DC location for deploying the service, i.e., d, a primary connectivity path p, and (optionally) a backup path b_p as protection to meet availability requirements. The selection process aims to minimize a resource consumption metric defined as:

$$c_{(j,d,p,b_p)} = \beta \cdot \frac{S_j}{M_d} + \alpha \cdot \left(\sum_{e \in \mathcal{E}_{j,d,p}} \frac{\eta_{j,e}}{W_e} + \sum_{e \in \mathcal{E}_{j,d,b_p}} \frac{\eta_{j,e}}{W_e} \right) \tag{5.1}$$

$c_{(j,d,p,b_p)}$ is the sum of two terms. The first term is related to the compute resources usage, and the second is the connectivity resources utilization (on the primary and backup paths). $\mathcal{E}_{j,d,p}$ and $\mathcal{E}_{j,d,b_p}$ are the set of links over paths p and b_p, respectively, between DC d and the AP. M_d is the total compute capacity of DC d, $\eta_{j,e}$ is the required number of wavelengths for the service over link e, and W_e is the overall wavelength capacity of link e. α and β are adjustable parameters that can be tuned based on available network resources, i.e., connectivity and compute resources, respectively.

The latency and availability models are explained in subsection 5.4.

The REP strategy, introduced in **Paper F**, is illustrated in Fig. 5.4 and works as follows. For each DC d in the set of DCs with enough compute capacity for service j ($\mathcal{D}$), REP looks for a path (g) between AP δ and DC

Table 5.1: Variable and parameters defined in Chapter 5.

γ: cost efficiency parameter	L_j: latency requirement
A_j: availability requirement	S_j: compute requirement
$c_{(j,d,p,b_p)}$: resource consumption	R_j: data rate requirement
p: primary path	b_p: backup path
d: arbitrary data center	δ: arbitrary access point
$\mathcal{E}_{j,d,p}$: links on path p between d and AP for service j	$\mathcal{E}_{j,d,b_p}$: links on path b_p between d and AP for service j
M_d: compute capacity of DC d	$\eta_{j,e}$: required number of wavelengths for service j over link e
W_e: wavelength capacity of link e	
α: tuning parameter for connectivity resources	β: tuning parameter for compute resources
$\mathcal{G}_{\delta,d}$: set of k-shortest paths between δ and d	g: an arbitrary path
	$\mathcal{Q}$: possible provisioning solutions
$\mathcal{P}_{\delta,d}$: primary paths between δ, d	$\mathcal{M}_{\delta,d,p}$: k-shortest paths between δ and d, and node disjoint with p
$\mathcal{B}_{\delta,d,p}$: backup path between δ, d and node disjoint with p	K_e: capacity of fiber link e
ρ_e: unavailability of link e	μ_n: unavailability of node n
N: number of nodes between AP and DC	ζ_e: length of link e
	h_j: holding time of service j
$\mathcal{X}_d$: set of deployed services in DC d during simulation time T	equ_life: equipment lifetime
	$\mathcal{I}$: set of DCs in the given network segment
$\mathcal{D}$: set of DCs with enough compute capacity for service j	$\mathcal{L}$: set of links in the given network segment
T: simulation time	
$\mathcal{Y}_e$: set of services using link e during simulation time T	$\mathcal{AC}$: set of accepted services after steady state, for duration equ_life

d. The path g is selected from $\mathcal{G}_{\delta,d}$, which is a pre-computed set of k-shortest paths between δ and d. If latency is not met, REP looks for the next candidate path in $\mathcal{G}_{\delta,d}$. Suppose the latency requirement over path g is met, but the availability is insufficient. In that case, g will be added to a set of possible

primary paths ($\mathcal{P}_{\delta,d}$) to check, in a later step, whether adding a backup path can help to meet availability requirements. If path g meets service latency and availability requirements and has enough connectivity resources, (d, g) will be added to the list of possible provisioning solutions, i.e., $\mathcal{Q}$.

After an option is added to $\mathcal{Q}$ or all paths in $\mathcal{G}_{\delta,d}$ are checked, REP considers paths in $\mathcal{P}_{\delta,d}$ to see if a backup path can be found. For each $p \in \mathcal{P}_{\delta,d}$, REP checks node-disjoint paths $b_p \in \mathcal{M}_{\delta,d,p}$, where $\mathcal{M}_{\delta,d,p}$ is pre-computed using the k-shortest path. If b_p meets the latency requirement and primary plus backup path $(p + b_p)$ meets the availability requirement, REP adds b_p to the list of possible backup path options ($\mathcal{B}_{\delta,d,p}$). After checking all paths in $\mathcal{M}_{\delta,d,p}$, REP validates whether p and $b_p \in \mathcal{B}_{\delta,d,p}$ have enough connectivity resources, and add (d, p, b_p) to the list of possible provisioning solutions ($\mathcal{Q}$). Finally, after considering all $d \in \mathcal{D}$, REP selects the option in $\mathcal{Q}$ that minimizes resource consumption metric $c_{(j,d,p,b_p)}$ as the ultimate provisioning solution for service j. If $\mathcal{Q}$ is an empty set, service j is rejected.

REP may accept more service requests than conventional provisioning strategies that do not add a backup path. However, the need for extra connectivity resources can impact the profitability of REP. In the following section, we explain a model to evaluate the profitability of REP.

5.4 Latency, Availability, and Profit Models

This section explains the latency and availability modeling to ensure service requirements are met. Also, the model for profit analysis of operators in a dynamic service provisioning scenario is presented.

Latency and availability modeling technique

A detailed explanation of availability and latency is provided in Sections 2.5 and 2.6. In this work, latency corresponds to the propagation and processing time between the user and the user plane function (UPF) using the formulation in Section 2.6. The latency is calculated as the sum of radio access network latency (i.e., baseband processing and over-the-air (OTA)), switching latency due to grooming at MN and/or ME, and propagation delay over the fiber links [43], [44].

The connection availability is a function of the availability of nodes and

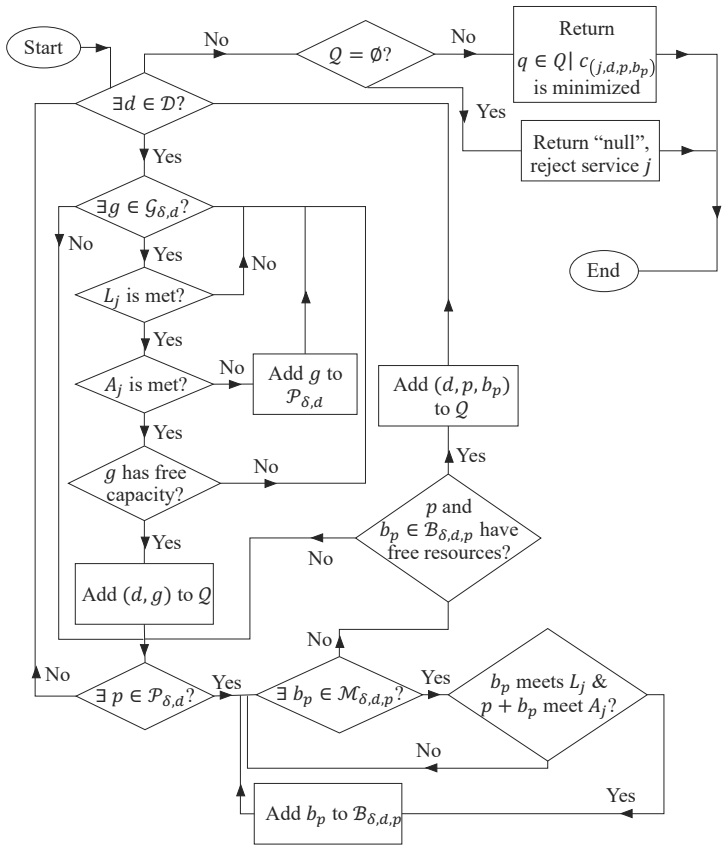


Figure 5.4: Flowchart of resource-efficient provisioning (REP) strategy.

links along the path from AP to DC defined by Eq. 2.2. Using the general formulation in Eq. 2.2, in this chapter, the availability is calculated as [44]:

$$a_{\mathrm{TN}} = \prod_{n=1}^{N} (1 - \mu_n) \times \prod_{e=1}^{N+1} (1 - \rho_e \times \zeta_e)\,, \tag{5.2}$$

where ρ_e and μ_n are the unavailability values of link e (i.e., per [km]) and node n (i.e., the probability that link e and node n do not work), respectively. ζ_e is the length of link e and N is the number of nodes between the AP and the DC

[44]. As shown in Eq. 2.3, in the presence of a backup path, the connection unavailability is a product of the unavailabilities of primary and backup paths (UA_p and UA_{b_p}). Then, the connection availability is defined as:

$$A_{total} = 1 - UA_p \times UA_{b_p} \tag{5.3}$$

Profit model

This subsection presents the model to evaluate the profitability of a dynamic service provisioning strategy.

In general terms, profit is the difference between generated revenue and cost. Operators get revenue by charging their customers for the provided services. The cost corresponds to the expenses related to the deployment cost (capital expenditure (CapEx)) and operational cost (operational expenditure (OpEx)) of the network and compute infrastructure. Accordingly, the total profit can be defined as:

$$Total\ Profit = \sum_j Profit_j = \sum_j Revenue_j - Cost_j \tag{5.4}$$

where $Revenue_j$ and $Cost_j$ are the revenue and cost associated with service j, respectively. We assume that an operator's total profit is the sum of profit associated with all accepted services during the network lifetime. We assume a cost-based pricing strategy. Thus, operators set a minimum price to charge their customers to cover their expenses [77].

The revenue generated by accepting service request j is the sum of the revenue due to assigning connectivity and compute resources for that service and is defined as:

$$Revenue_j = Connectivity\ Revenue_j + Compute\ Revenue_j \tag{5.5}$$

The revenue resulting from providing connectivity resources for service j is defined as:

$$Connectivity\ Revenue_j = conn\_char \times surcharge \times h_j \times R_j \tag{5.6}$$

where *conn_char* is the *connectivity charge*, which is the cost of provisioning one data unit and set by the operator (i.e., measured in monetary unit [MU] over data unit [DU]). The operator specifies the *surcharge* value to have enough revenue margin. By increasing the *surcharge*, for the same number of accepted service requests, the revenue will increase. h_j is the holding time (i.e., measured in time unit [TU]), and R_j is the data rate (i.e., measured in [DU] over [TU]) of the service j.

The revenue generated by providing compute resources for service j is calculated as:

$$Compute\ Revenue_j = comp\_char \times surcharge \times h_j \times S_j \tag{5.7}$$

where *comp_char* is the *compute charge*, which is the price for requesting one compute unit per [TU] (measured in [MU] over compute unit [CompU] over [TU]), which is decided by the operator. *surcharge* and h_j values are the same, as explained in Eq. 5.6. S_j is the amount of requested compute resources (measured in [CompU]) by service j.

The $Cost_j$ in Eq. 5.4 is the cost of providing service j. It is the expenditure for deploying (CapEx) and maintaining (OpEx) connectivity and compute resources. The cost due to accepting service request j is the summation of the cost of connectivity and compute resources required by the service and is defined as:

$$Cost_j = Connectivity\ Cost_j + Compute\ Cost_j \tag{5.8}$$

The connectivity cost of service j is derived as:

$$Connectivity\ Cost_j = \frac{h_j}{equ\_life} \times \left(\sum_{\mathcal{TX}} TxRx\_usage \times TxRx\_cost + \sum_{\mathcal{E}} link\_usage \times fiber\_cost \right) \tag{5.9}$$

where $\mathcal{TX}$ and $\mathcal{E}$ are the sets of transceivers (TxRx) and links along the path

from AP to DC, respectively. The $TxRx\_usage$ shows the utilization of a transceiver and can be obtained by $\frac{R_j}{\text{TxRx transmission rate}}$. Likewise, $link\_usage$ shows the utilization of a link and can be defined by $\frac{R_j}{\text{fiber link capacity}}$. The $TxRx\_cost$ and $fiber\_cost$ consider the CapEx and OpEx of transceiver and fiber, respectively. equ_life is the average equipment lifetime. The intuition behind normalizing cost to equipment lifetime is that the equipment is installed once but will be used to provide many services over time.

Following the same logic, the compute cost to provision service j is calculated as:

$$Compute\ Cost_j = \frac{h_j}{equ\_life} \times DC\_usage \times DC\_cost \tag{5.10}$$

where DC_usage measures used compute resources for provisioning service j out of the selected DC capacity (i.e., $\frac{S_j}{\text{DC capacity}}$). DC_cost is the cost of the selected DC.

5.5 Performance Evaluation

This section first presents the assumptions on the service requirements. Then, the performance of the proposed strategy in terms of service blocking ratio, resource utilization, and profit is discussed. We developed an ad-hoc, Python-based, event-driven simulator to evaluate our strategy.

We consider three service categories (i.e., media streaming (MS), massive Internet-of-things (mIoT), and 6 degrees of freedom (6-DoF) virtual reality (VR)) with different latency, availability, data rate, and compute requirements shown in Table 5.2. The number of users for each service request is chosen randomly with uniform distribution within the range specified in Table 5.2. The range values should be multiplied by 100 to get the number of users (it is shown this way due to space constraints).

The percentage of different types of service requests is expected to change in the future. Particularly, 6-DoF VR is expected to replace traditional video streaming [78], and we will have more services with strict availability requirements in the future. Likewise, operators expect the number of mIoT devices to increase. To consider this trend, we evaluate three scenarios: current (i.e., today), short-term (i.e., three years from today), and long-term (i.e., five years from today). The expected composition of services in each scenario is

presented in Table 5.2.

Table 5.2: Services types and their (per user) requirements, range of users for each service type, and traffic composition [78], [79].

		MS	mIoT	6-DoF VR
Requirements	Latency [ms]	4000	5	10
	Availability [%]	99.99	99.999	99.99
	Data Rate [Mbps]	20	0.1	500
	Compute [CompU]	0.002	0.005	0.2
Traffic composition	Number of users (must multiply by 100)	$(2, 10)$	$(80, 150)$	$(0.1, 0.3)$
	Today [%]	70	25	5
	3-years [%]	30	50	20
	5-years [%]	10	60	30

The detailed assumptions on the number of nodes, rings, the fiber link capacity, the DCs compute capacity (i.e., in access, metro, and core), the cost of connectivity and compute resources, and latency contribution of different elements to the total latency are presented in **Paper G**.

We assume that the service arrivals follow a Poisson process in which the inter-arrival time is exponentially distributed with rate λ. The mean value of the holding time is 24 [TU], where one [TU] is one hour.

We evaluate REP by comparing its performance to a conventional service provisioning strategy, referred to as no path protection (NPP). NPP selects a DC and a connectivity path between the AP and the DC, but it does not add any backup path. We assume that the connectivity and compute weighting factors (i.e., α and β) in Eq. 5.1 are 0.5 and 1, respectively.

Figure 5.5 shows the service blocking ratio of REP compared to NPP as a function of network load for three considered scenarios, i.e., today (T), 3 years (3Y), and 5 years (5Y) ahead. The blocking ratio represents the proportion of rejected service requests relative to the total number of received service requests (in percentage). Fig. 5.5 illustrates that REP has a lower blocking ratio compared to NPP. The blocking ratio gain in REP is achieved through the possibility of adding a backup path. The lower blocking ratio of

REP compared to NPP is more evident in the 5-years (5Y) scenario than in the 3-years (3Y) case, which, in turn, is more pronounced than today's (T) scenario. Indeed, the percentage of services with stringent availability requirements increases with time. By providing backup connectivity resources, REP can deploy services in more central locations of the network, thus, addressing the compute resources limitation in the access segment. The figure underlines the importance of benefiting from centralized and more abundant compute resources using a backup-enhanced compute off-loading strategy (e.g., REP) to support next-generation services.

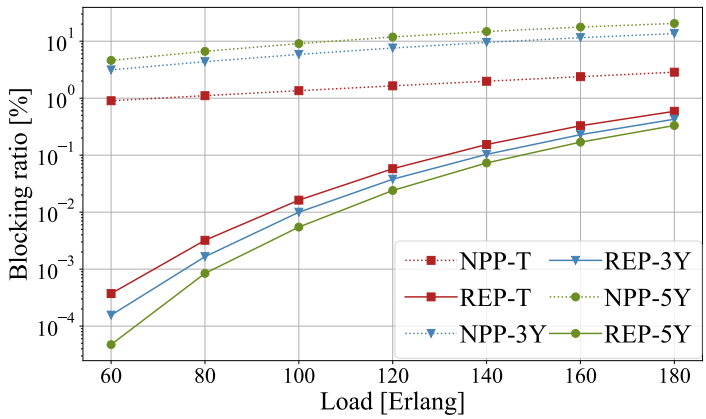


Figure 5.5: Service blocking ratio as a function of the network load. Three traffic scenarios: today (T), 3-years (3Y), and 5-years (5Y). Reprinted from **Paper G**.

To understand the impact of extra backup resources, we evaluate the average connectivity and compute resource utilization of REP compared to NPP. The average compute resources utilization (AVCU) of DCs in a given network segment (i.e., access, metro, or core) is defined as:

$$AVCU = \frac{1}{|\mathcal{I}|} \sum_{d \in \mathcal{I}} \sum_{j \in \mathcal{X}_d} \frac{h_j \times S_j}{T \times M_d} \tag{5.11}$$

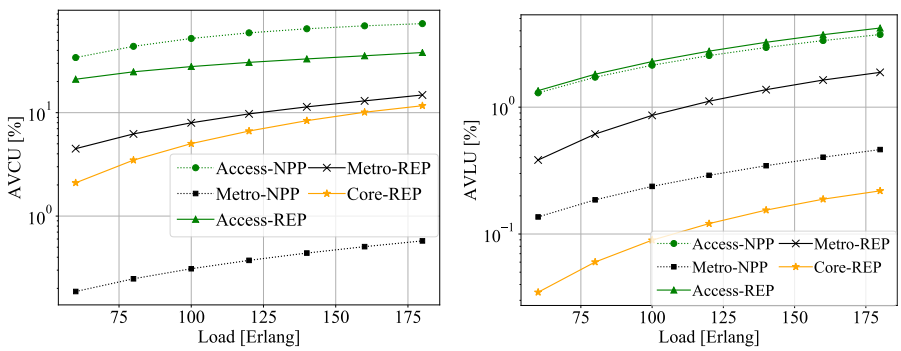
where $\mathcal{I}$ is the set of DCs in a given network segment, $\mathcal{X}_d$ is the set of deployed services in DC d during the simulation time T, h_j is the holding time of

service j, and S_j is the compute requirement of service j. Likewise, the average connectivity resources utilization (AVLU) in a given network segment is calculated as:

$$AVLU = \frac{1}{|\mathcal{L}|} \sum_{e \in \mathcal{L}} \sum_{j \in \mathcal{Y}_e} \frac{h_j \times R_j}{T \times K_e} \tag{5.12}$$

where $\mathcal{L}$ is the set of links in a given network segment, $\mathcal{Y}_e$ is the set of services using link e during the simulation time T, K_e is the capacity of fiber link e, and R_j is the data rate of service j.

Figures 5.6a and 5.6b show AVCU and AVLU of REP compared to NPP, respectively, where the service requirements are as described in Table 5.2. Figure 5.6a shows that, by adopting REP, the AVCU of access DCs has decreased while it is increased for metro and core DCs thanks to a higher possibility of off-loading in REP compared to NPP. However, Fig. 5.6b shows the AVLU of both REP and NPP are very similar, and the lower blocking ratio of REP (Fig. 5.5) is achieved at the cost of slightly higher connectivity resource usage.

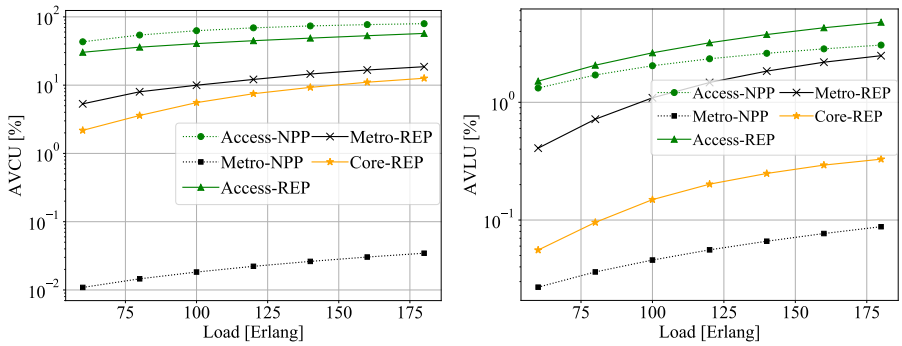


(a) Average compute utilization. **(b)** Average connectivity utilization.

Figure 5.6: Connectivity and compute resource utilization vs. load for the long-term scenario (5-years). 6-DoF VR has availability of 99.99% and compute requirement of 0.2 [CompU] per user.

Figures 5.7a and 5.7b show AVCU and AVLU of REP compared to NPP, respectively, where the availability and compute resources requirements of 6-DoF VR are assumed to be more stringent (i.e., availability of 99.999% and

compute resources requirement of 2 [CompU] per user) to evaluate further the performance of REP in supporting next-generation scenarios. The other service requirements are assumed to be the same as Table 5.2. We observe similar trends to Figures 5.6a and 5.6b. REP can utilize compute resources in the metro and core network segments, thus, off-loading the compute resources in access DCs (Fig. 5.7a). This results in a slight increase in connectivity resource utilization (Fig. 5.7b) but offers a significant improvement in blocking ratio performance compared to NPP.



(a) Average compute utilization.

(b) Average connectivity utilization.

Figure 5.7: Connectivity and compute resource utilization vs. load for the long-term scenario (5-years). 6-DoF VR has availability of 99.999% and compute requirement of 2 [CompU] per user.

To better understand the impact of a lower blocking ratio in REP and the way it utilizes connectivity and compute resources, we compare the profitability of REP and NPP. To evaluate the profit, we consider a *surcharge* value of 1.2, which translates into a 20% revenue margin, and an equipment lifetime (equ_life) of 5 years.

As already explained in the profit model, *conn_char* in Eq. 5.6 is the cost of provisioning one data unit (DU), and *comp_char* in Eq. 5.7 is the price of provisioning one compute unit (CompU) per time unit (TU) charged by the operator. Operators need to properly set the *conn_char* and *comp_char* to charge their customers for a minimum price so that they can return their investment and at least cover CapEx and OpEx [77].

One approach to determine the values of *conn_char* and *comp_char* is to calculate the average cost of provisioning services before actual network

operations. This can be accomplished by simulating a service provisioning scenario in advance, prior to the start of actual provisioning. During this simulation, we must wait until the network reaches a steady state, indicated by a fixed blocking ratio. Following this, network operations will continue for the duration equal to the equipment lifetime (equ_life). We calculate the average cost of accepted services during equ_life (i.e., set $\mathcal{AC}$) to find $conn\_char$ and $comp\_char$. Following the intuition, we calculate $conn\_char$ as:

$$conn\_char = \frac{1}{equ\_life \times |\mathcal{AC}|} \times \sum_{j \in \mathcal{AC}} \frac{\sum_{\mathcal{TX}} TxRx\_usage \times TxRx\_cost + \sum_{\mathcal{E}} link\_usage \times fiber\_cost}{R_j} \tag{5.13}$$

All elements in Eq. 5.13 are already explained in Eq. 5.6 and 5.9. Using this intuition, the $conn\_char$ for today, 3-years, and 5-years scenarios are 2.6×10^{-12} [MU/DU], 4×10^{-12} [MU/DU], and 4.6×10^{-12} [MU/DU], respectively, where one [DU] is 1 [Mbit]. One MU corresponds to the cost of a 10 [Gbps] Tx/Rx.

Following the same procedure, $comp\_char$ is calculated as:

$$comp\_char = \frac{1}{equ\_life \times |\mathcal{AC}|} \sum_{j \in \mathcal{AC}} \frac{DC\_usage \times DC\_cost}{S_j} \tag{5.14}$$

All the elements in this equation are already explained in Eq. 5.7 and 5.10. The values of the $comp\_char$ for today, 3-years, and 5-years scenarios are 2.4×10^{-6} [MU/(TU·CompU)], 2.1×10^{-6} [MU/(TU·CompU)], and 2×10^{-6} [MU/(TU·CompU)], respectively, where one [CompU] corresponds to 1 CPU-core.

Figure 5.8 shows the profit ratio of REP compared to NPP as a function of load. The service requirements are presented in Table 5.2. Adopting REP strategy brings profit gain for all load values. Indeed, the possibility of adding a backup path in REP results in accepting more service requests and generating higher revenues than NPP. Although adding a backup path comes with

a slight increase in the cost of connectivity resources, its impact on the profitability of REP is small. Accordingly, when the number of services with stringent availability requirements increases (i.e., in the 5-years scenario), REP can show significant profit gains. Following the same reasoning, the profit gain of the 3-years ahead is larger than today's scenario.

Figure 5.9 shows the profit ratio of REP compared to NPP for a more severe case where the availability requirement of 6-DoF VR is 99.999%, and its compute requirements are 2 [CompU] per user, while the requirements of the other services are the same as in Table 5.2. This figure illustrates that even when the service requirements are very stringent, REP brings significant profit gains compared to NPP. REP can achieve this gain by accepting more service requests, bringing larger revenues, and compensating for the extra cost of connectivity resources.

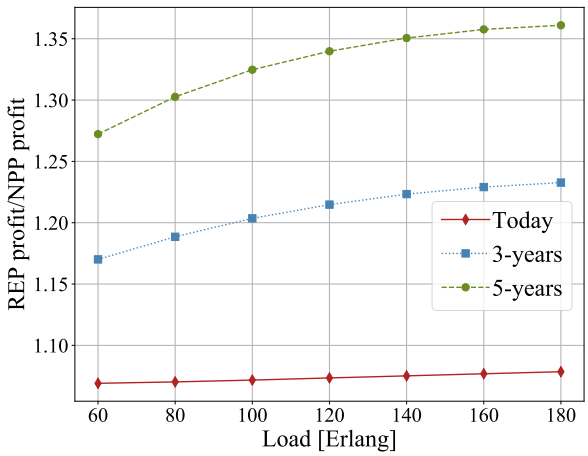


Figure 5.8: Profit gain as a function of load. 6-DoF VR requirement is 99.99% for availability and 0.2 [CompU] per user for compute resources. Reprinted from **Paper G**.

5.6 Summary

This chapter first presents a simple scenario to show the cost benefits of centralized service deployment by adding a backup path while meeting strict ser-

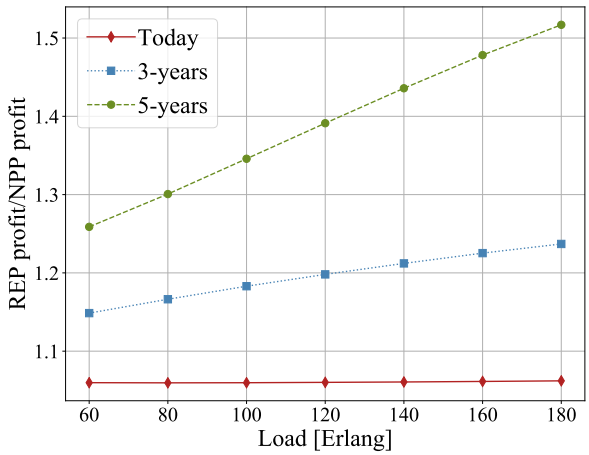


Figure 5.9: Profit gain as a function of load. 6-DoF VR requirement is 99.999% for availability and 2 [CompU] per user for compute resources. Reprinted from **Paper G**.

vice requirements. Deploying a backup path uses extra connectivity resources. However, the benefits of using large and centralized DCs (i.e., abundant compute resources, lower cost because of the economy of scale, and multiplexing gain of high tier TN segments), result in up to 74% savings of infrastructure cost.

After that, this chapter extends the evaluation of centralized deployment in dynamic provisioning of 5G and beyond services with stringent requirements (i.e., a long-term scenario). Compute off-loading to central DCs (e.g., located in metro and core segments) has many-fold advantages. These DCs have abundant compute resources, and the cost of service processing in central DCs is lower than the edge DCs. However, the strict service latency and availability requirements limit the distance that can be traversed to reach metro and core DCs. The only solution to address the demanding latency constraint is to deploy services close to the user (e.g., in the access segment). To meet availability requirements, a backup path can be added. On the other hand, the price of additional connectivity resources may limit the cost benefits of centralized deployment and operators' profit.

REP is our proposed method which is a backup-enhanced compute off-loading strategy. To meet service availability requirements, REP benefits

from the possibility of adding a backup path (when needed). This, in turn, leads to off-loading the compute resources at the access DCs to more central DCs. Accordingly, access DCs will be available for services with very stringent requirements that cannot be deployed in central locations of the network.

We evaluate the service blocking ratio, resource utilization, and profit of adopting REP compared to a conventional provisioning strategy that does not add a backup path. We consider today, three years ahead, and five years ahead scenarios. It is shown that REP can maximize resource efficiency in a dynamic service provisioning scenario such that service latency and availability requirements are met. Simulation results show that REP brings a lower blocking ratio and higher profit gains because it accepts more services with stringent availability requirements and addresses the compute resource limitations in the access segment. These gains are more evident in future scenarios (e.g., in 5 years from now), as more services with strict requirements need to be provided compared to today's scenario.

CHAPTER 6

Summary of included papers

This chapter provides a summary of the papers included in the thesis. Full versions of the papers are appended in Part II. The layout of the papers has been revised to be consistent with the layout of the thesis.

6.1 Paper A

Maryam Lashgari, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti
Fiber- vs. Microwave-based 5G Transport: a Total Cost of Ownership Analysis
European Conference on Optical Communication (ECOC), Basel, Switzerland, Sep. 2022.

Operators are interested in a cost-efficient transport network (TN) deployment solution for 5G networks. This paper presents a TCO analysis of three 5G transport architectures based on fiber and microwave technologies for high layer functional split option. It also considers three different network deployments related to the initial stage deployment of 5G in dense urban, urban,

and sub-urban areas. Simulation results show that the TCO of the microwave-based solution is lower than the fiber-based ones in all considered urban areas. The TCO gains depend on the area under exam and the cost of fiber trenching and microwave equipment.

6.2 Paper B

Maryam Lashgari, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti

Techno-economics of Fiber vs. Microwave for Mobile Transport Network Deployments (Invited)

To be published in Journal of Optical Communications and Networking (JOCN), vol. 15, no. 7. ©Optica Publishing Group. Reprinted with permission.

DOI: 10.1364/JOCN.482865.

Deploying a TN solution which is cost-efficient and can meet the capacity, latency, and reliability requirements of 5G services is vital for operators. This paper presents fiber- and microwave-based TN architectures for high layer and low layer functional split options. The TCO, latency, and reliability performance of different deployment options are evaluated. Moreover, this paper investigates the impact of using equipment with reconfigurability capabilities in high-layer split option architectures. The results indicate that in most of the considered scenarios (high layer or low layer split option in dense urban, urban, and sub-urban areas), a microwave-based TN exhibits lower TCO than a fiber-based architecture. On the other hand, the TCO gain varies with the type of urban area, reconfigurability features, selected functional split option, and the cost of fiber trenching and microwave equipment (which can differ depending on the country/operator). In particular, the fiber- and microwave-based solutions have comparable TCO for low layer functional split option in a dense urban area, where the average link length is relatively short. The architectures with reconfigurability capabilities have higher TCO than their counterpart without such a feature. However, even with the reconfigurability feature, the microwave-based alternative has lower TCO than the fiber-based one in all urban areas for the high layer functional split option. Finally, to evaluate latency performance, the requirements of eMBB and URLLC services are considered. The investigated fiber and microwave solutions have almost

similar average latency and can meet the requirements of 5G and beyond 5G services. The connection availability performance of considered architectures is almost similar and within the range required by 3GPP. In very latency-critical scenarios (i.e., where the latency requirement of low layer split option is 0.025 [ms]), a small number of cells cannot meet the requirements using microwave-based architecture, mainly because of the need for multiple microwave hops.

6.3 Paper C

Maryam Lashgari, Federico Tonini, Massimiliano Capacchione, Lena Wosinska, Gabriele Rigamonti, and Paolo Monti
Techno-economics of 5G Transport Deployments
Proc. of Next-Generation Optical Communication: Components, Sub-Systems, and Systems XII, edited by Guifang Li, Kazuhide Nakajima, Atul K. Srivastava, SPIE 12429, San Francisco, California, United States, Jan. 2023.
DOI: 10.1117/12.2652618.

The microwave deployment solution in **Paper B** could not satisfy critical latency requirements in all sites for the low layer split option. **Paper C** presents an additional hybrid fiber-microwave architecture intending to meet very stringent latency requirements. We evaluate the hybrid architecture regarding the TCO and latency performance. It can meet the stringent latency requirements for the low layer split option (0.025 [ms]) in 100% of the cell sites. The hybrid architecture uses microwave links in most sites to reduce costs compared to the fiber-only solution. It introduces fiber where the multiple microwave hops violate stringent latency requirements.

6.4 Paper D

Maryam Lashgari, Lena Wosinska, and Paolo Monti
A Shared-Path Shared-Compute Planning Strategy for a Resilient Hybrid C-RAN
21st International Conference on Transparent Optical Networks (ICTON), Angers, France, July 2019.

ska and P. Monti, "A Shared-Path Shared-Compute Planning Strategy for a Resilient Hybrid C-RAN," 21st International Conference on Transparent Optical Networks (ICTON), Angers, France, 2019, pp. 1-6. DOI: 10.1109/ICTON.2019.8840404. .

Designing a resilient network for uninterrupted operation in 5G networks is crucial. The paper proposes a strategy to guarantee the survivability of services in the presence of a single failure in the cloud data centers or in the nodes/links in the midhaul segment. The proposed strategy is referred to as shared-path shared-compute planning (SPSCP). It assigns two node-disjoint connectivity paths to connect each radio aggregation unit (RAU) (i.e., nodes to which cell sites are connected) to primary and backup cloud data centers. The SPSCP strategy aims at minimizing the overall design cost by maximizing sharing of the backup connectivity and compute resources among RAUs. SPSCP uses a shareability metric when selecting cloud data centers and connectivity paths between RAUs and the data centers to increase the possibility of sharing. Results indicate that the SPSCP strategy can offer 28% cost savings compared to a benchmark strategy that assigns dedicated backup resources. Moreover, the cost savings are 23% compared to another benchmark that shares backup resources as much as possible in the second stage after deploying data centers.

6.5 Paper E

Maryam Lashgari, Carlos Natalino, Luis M. Contreras, Lena Wosinska, and Paolo Monti
Cost Benefits of Centralizing Service Processing in 5G Network Infrastructures
Asia Communications and Photonics (ACP) Conference, Chengdu, China, Nov. 2019.

Operators seek network design solutions to minimize the overall infrastructure cost. This paper considers adding a backup path to meet the service availability requirements of services. It investigates the trade-offs between the cost benefits of centralizing service deployment in a few large DCs and the cost of extra connectivity resources on the backup path. The network is deployed to provision four 5G services with different latency and availability

requirements. It is found that the economy of scale benefits of centralized deployment (which is enabled by adding a backup path) yields up to 74% savings in overall infrastructure cost.

6.6 Paper F

Maryam Lashgari, Lena Wosinska, and Paolo Monti
End-to-End Provisioning of Latency and Availability Constrained 5G Services
IEEE Communications Letters, vol. 25, no. 6, pp. 1857-1861, June 2021.

DOI: 10.1109/LCOMM.2021.3063262 .

This paper proposes a strategy for the dynamic provisioning of 5G services with strict latency and availability requirements. The proposed strategy aims at maximizing the efficiency of connectivity and compute resources by encouraging centralized service processing (i.e., compute off-loading to large DCs). Accordingly, it can leverage multiplexing gains in high-tier transport network segments (e.g., metro and core). Moreover, it has access to abundant compute resources available in the large DCs (e.g., metro and core) compared to access DCs. The strategy, i.e., referred to as resource-efficient provisioning (REP), selects a DC and a connectivity path for each service request by measuring compute and connectivity resource utilization of each option. Moreover, REP can add a backup path to not violate the service availability requirements when deploying services in the central DC locations. Hence, REP is a backup-enhanced compute off-loading strategy. The performance of REP is evaluated against a conventional service provisioning strategy that does not add backup connectivity resources (leading to the need for deploying more services in small DCs at the network edge). Results indicate that REP can improve the service blocking ratio by up to four orders of magnitude compared to the conventional approach. This improvement is up to two orders of magnitude considering another 5G use case with relaxed service latency and availability requirements.

6.7 Paper G

Maryam Lashgari, Federico Tonini, Lena Wosinska, Luis M. Contreras, and Paolo Monti
Next-Generation Service Deployment with Compute Off-Loading: a Profit Analysis Perspective
Submitted to IEEE Network in Apr. 2023.

One of the challenges for operators is finding a resource-efficient service provisioning approach to meet stringent latency, availability, compute, and connectivity requirements. This paper presents a profit analysis of compute off-loading strategies in a dynamic service provisioning scenario. The study provides a guideline for operators to understand the advantages of using compute off-loading strategies to provide next-generation services. Using a compute off-loading strategy can preserve the limited compute resources at the edge DCs for services with very stringent latency requirements, where propagation delay prevents deploying those services at distant locations. This paper considers three representative scenarios with various compositions of services with different requirements regarding latency, availability, compute, and data rate. The considered scenarios refer to the existing traffic in the network today, in the short-term (i.e., three years from now), and in the long-term (i.e., five years from now). REP is proposed as a backup-enhanced compute off-loading method, which considers adding a protection path (when needed) for centralized service deployment. The service blocking ratio, connectivity and compute resource utilization, and profitability of REP are assessed compared to a conventional service provisioning approach. Simulation results show that REP brings higher profit gain and lower service blocking ratio than a conventional approach (which does not add any backup path) in all considered scenarios. Moreover, these gains are increasing in time, i.e., they are the highest in the 5-year from now scenario, where many service requests with very stringent requirements must be provisioned.

CHAPTER 7

Concluding Remarks and Future Work

This chapter summarizes the main contributions of the thesis and draws some concluding remarks. Then, it highlights some directions for future work.

7.1 Conclusions

The stringent service requirements in 5G and beyond can be met by network densification, i.e., deploying new cell sites in the network close to the users. The new cell sites must be connected to the mobile core network through a suitable TN option. Operators must carefully evaluate different technologies, functional split options, and reconfigurability levels to identify the most cost-effective TN solution. Moreover, achieving low latency and high availability for the provisioned services is critical. To ensure a reliable TN, resilient design strategies should also be employed to provide survivability against failures while minimizing overall infrastructure costs. Once a TN is deployed, operators must dynamically provision services with different requirements. Operators aim to utilize efficiently available resources and maximize their profit in service provisioning. Overall, the successful deployment and operation of 5G and beyond networks require careful consideration of all these key factors.

This thesis proposes and evaluates design and service provisioning strategies. The aim is to maximize resource efficiency and meet the stringent service requirements of 5G and beyond services.

Several TN architectures based on fiber and microwave technologies using high layer split (HLS) and low layer split (LLS) options with different reconfigurability capabilities are analyzed. A comprehensive framework is proposed to evaluate the TCO, latency, and availability performance of the investigated architectures. The dense urban, urban, and sub-urban areas are considered. Network dimensioning and components' costs are based on real data provided by a system vendor. Results show that the TCO performance gains of microwave deployment compared to their fiber counterpart vary depending on the chosen functional split option, the specific deployment area under exam, and the cost of fiber trenching and microwave equipment. The impact of these aspects is more pronounced in dense urban areas than in urban and sub-urban. In particular, using an LLS option, the TCO of fiber and microwave-based solutions are comparable in dense urban. On the other hand, the longer link length in urban and sub-urban areas leads to high fiber deployment cost, and consequently, microwave becomes more cost-efficient. The considered TN deployment options can meet user plane latency requirements of URLLC and eMBB services in the case of HLS option. On the other hand, in the LLS option, the microwave-based solution can meet the requirements of eMBB and URLLC-T in all sites. However, in the case of LLS and URLLC-S (with more stringent latency requirements), only single-hop microwave links can meet the latency requirements. Accordingly, a fiber-only or a hybrid fiber-microwave alternative must be used to meet the latency requirement in all sites. The investigated architectures have similar connection availability performance within the range required by 3GPP.

In TN deployment, ensuring the survivability of services running on the network in the presence of failures is essential. The resilient TN must be designed while minimizing overall infrastructure costs. A resource-efficient strategy is proposed (referred to as shared-path shared-compute planning (SPSCP)) to provide resiliency against a single failure in DCs (where 5G protocol stack functions are processed) or in the nodes/links of the midhaul segment. SPSCP assigns a primary and a backup DC to each aggregation node (i.e., nodes to which cell sites are connected). The aggregation node is connected to its backup DC through a backup path, which is node-disjoint with the primary

path. SPSCP tries to maximize sharing among backup resources to reduce the total infrastructure costs. To achieve this, it considers the shareability potential when assigning primary and backup DCs and connectivity paths. The performance of SPSCP is evaluated against three benchmarks. Our findings show that the cost-efficiency of SPSCP can result in 28% savings compared to a benchmark that does not share resources and relies only on dedicated backup resources. The cost savings of SPSCP is 23% compared to another benchmark that only supports simple sharing. This benchmark tries to modify the pairing between aggregation nodes and their backup DCs at a second stage (after locations of DCs and connectivity paths are specified) to improve sharing. Instead, our strategy considers the potential for sharing backup resources while deploying DCs and connectivity paths.

Regarding dynamic service provisioning, operators are interested in guidelines to use their infrastructure resources efficiently while maximizing their profit in various scenarios. The compute off-loading strategies can leverage abundant compute resources at central DCs for deploying services when the latency requirements are not very strict. It makes compute resources in the edge DCs available for services with stringent latency requirements. Hence, this thesis presents a strategy, referred to as resource-efficient provisioning (REP), which is a *backup-enhanced compute off-loading* method to maximize resource efficiency. REP selects a DC as central as possible and a connectivity path while meeting the latency and availability requirements. Moreover, REP can add a backup path (where needed) to satisfy availability requirements. We considered beyond 5G services with different latency, availability, compute, and data rate requirements. The performance of REP is evaluated in terms of service blocking ratio, resource utilization, and profit in three scenarios, i.e., today, short-term (3 years from now), and long-term (5 years from now). Simulation results show that REP has a lower service blocking ratio (up to four orders of magnitude) and higher profit (up to 35% to 50%) compared to a conventional approach that does not add a backup path. Using REP brings higher gains in terms of profit and blocking ratio in future scenarios (e.g., in 5 years) when the percentage of services with strict requirements is more evident in the network.

7.2 Future Work

In terms of future work, it would be interesting to consider more architectural options for the TN deployment. In particular, given the challenges of multi-hop microwave links, we explored a hybrid fiber-microwave architecture to enhance latency performance. Although the architecture was effective, other hybrid architectures are worth further investigation. For example, one possibility is to eliminate the networking devices in the hybrid architecture, which may decrease the TCO and improve latency performance. However, this approach may compromise the network's scalability. Another option is to reduce the use of fibers in the hybrid architecture by leveraging microwave technology for all last-hop links. Implementing this approach may result in a reduced TCO compared to the current hybrid architecture, bringing it more in line with the TCO of a microwave-based architecture.

With the increasing penetration of critical communication services and applications with demanding availability requirements, the resiliency of mobile networks is an essential element for operators. For example, the reliability requirement of 6G applications can be in the order of 99.9999999% [28], [31]. One approach for enhancing the reliability performance of our considered TN architectures is to provide backup connectivity paths by integrating several transmission technologies. One way is to use microwave links as a backup path for fiber links. It is not required to deploy this backup solution broadly in the network but only in the hotspots where ultra-reliable services are delivered. It would be interesting to investigate the impact of adding such a backup path on the TCO and availability performance of TN deployment options.

Utilizing connectivity and compute resources efficiently in the network becomes even more crucial with the advent of new resource-hungry applications such as 6-DoF VR. Therefore, finding more resource-efficient service provisioning strategies is an important step toward future network operations. The machine learning algorithms can be used to improve the performance of our backup-enhanced compute off-loading strategy (i.e., REP).

Various machine learning techniques can improve REP in different ways. One approach is to use supervised learning and predict future traffic. Based on traffic predictions, some service requests can be proactively rejected to reserve the capacity for more critical services or those that generate higher revenue. Moreover, traffic prediction can be leveraged to decide where to deploy services to avoid future bottlenecks in the network. Accordingly, over-

all service blocking ratio performance can be improved. Another interesting direction would be to use reinforcement learning (RL) to learn the system's behavior. RL can aid in service scheduling by determining whether to accept or reject a service request. Additionally, RL can be employed for service provisioning, enabling decisions such as whether or not to add a protection path, which connectivity and compute resources to use, and where to deploy the service to achieve optimal resource efficiency. Both approaches have pros and cons. The approach based on supervised learning and traffic prediction is less complex but relies on prediction accuracy. On the other hand, RL does not need past data about the network. Instead, it interacts with the system to learn its characteristics. However, finding a suitable reward function for RL can be challenging.

A possible approach to further improve resource utilization is migrating accepted services and re-provisioning them using alternative DCs or paths. Adding a service migration strategy can extend our proposed method (REP). For this purpose, using a supervised approach to predict traffic would be interesting. Then, an algorithm must determine when the reconfiguration is triggered, which services are migrated, and where the services will be deployed. The reconfiguration process can be initiated, e.g., when the service blocking ratio exceeds a certain threshold or when the resource utilization of given DCs surpasses a particular value. The results may change depending on the selected threshold and the algorithm for migration. Another possibility is to use the RL approach for service migration. Thus, RL agent can choose a suitable threshold, the services to migrate, and the new placement of services. Finding a proper reward function to consider all these aspects while minimizing the service blocking ratio is challenging.

References

[1] ITU, *Setting the scene for 5G: Opportunities & challenges*, `https://www.itu.int/en/ITU-D/Documents/ITU_5G_REPORT-2018.pdf`, 2018.

[2] Q. Chen, J. Wang, and H. Jiang, *URLLC and eMBB coexistence in MIMO non-orthogonal multiple access systems*, arXiv:2109.05725, 2021.

[3] A. Shahraki, M. Abbasi, M. J. Piran, and A. Taherkordi, *A comprehensive survey on 6G networks: Applications, core services, enabling technologies, and future challenges*, arXiv:2101.12475v2, 2021.

[4] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6G: A comprehensive survey," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 334–366, 2021.

[5] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5G wireless networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.

[6] 3GPP, "TR 38.801, study on new radio access technology: Radio access architecture and interfaces," Technical report, version 14.0.0, Mar. 2017.

[7] L. M. P. 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, 2019.

[8] J. Lun, D. Grace, A. Burr, Y. Han, K. Leppanen, and T. Cai, "Millimetre wave backhaul/fronthaul deployments for ultra-dense outdoor small cells," in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2016, pp. 187–192.

[9] 3GPP, "TS 38.470, F1 general aspects and principles," Technical specification, version 17.0.0, Apr. 2022.

[10] O-RAN Open Xhaul Transport Working Group 9, "Xhaul transport requirements," Technical specification, version 1.00, Feb. 2021.

[11] O-RAN Open Fronthaul Interfaces Working Group 4, "Control, user and synchronization plane specification," Technical specification, version 7.01, Apr. 2022.

[12] T. Naveh, "Mobile backhaul: Fiber vs. microwave," *Ceragon White Paper*, vol. 1, pp. 1–11, Oct. 2009.

[13] Ericsson, "Ericsson microwave outlook," Report, Oct. 2021.

[14] J. Saunders and N. Marshall, "Mobile backhaul options spectrum analysis and recommendations," ABI Research, Research Report, Sep. 2018.

[15] China Mobile Research Institute; Alcatel-Lucent; Nokia; ZTE; Broadcom; Intel, *Next Generation Fronthaul Interface*, `https://docplayer.net/46899964-White-paper-of-next-generation-fronthaul-interface.html`, White paper, Jun. 2015.

[16] J. M. Simmons, *Optical Network Design and Planning*. Springer International Publishing, 2014.

[17] B. Naudts, M. Kind, F.-J. Westphal, S. Verbrugge, D. Colle, and M. Pickavet, "Techno-economic analysis of software defined networking as architecture for the virtualization of a mobile network," in *European Workshop on Software Defined Networking*, 2012, pp. 67–72.

[18] F. Marzouk, M. Lashgari, J. P. Barraca, *et al.*, "Virtual networking for lowering cost of ownership," in *Enabling 6G Mobile Networks*, J. Rodriguez, C. Verikoukis, J. S. Vardakas, and N. Passas, Eds. Cham: Springer International Publishing, 2022, pp. 331–369.

[19] M. R. Raza, M. Fiorani, A. Rostami, P. Öhlen, L. Wosinska, and P. Monti, "Dynamic slicing approach for multi-tenant 5G transport networks [invited]," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 10, no. 1, A77–A90, 2018.

[20] Next Generation Mobile Network (NGMN) Alliance, "5G white paper 2," White paper, version 1.0, Jul. 2020.

[21] Next Generation Mobile Network (NGMN) Alliance, "5G E2E technology to support verticals URLLC requirements," Final Deliverable, version 1.6, Feb. 2020.

[22] L. M. P. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, "Deployment guidelines for cloud-RAN in future mobile networks," in *IEEE 11th International Conference on Cloud Networking (CloudNet)*, 2022, pp. 141–149.

[23] A. Greenberg *et al.*, "The cost of a cloud: Research problems in data center networks," *CCR*, 2008, DOI: 10.1145/1496091.1496103.

[24] P. Öhlén, B. Skubic, A. Rostami, *et al.*, "Data plane and control architectures for 5G transport networks," *Journal of Lightwave Technology*, vol. 34, no. 6, pp. 1501–1508, 2016.

[25] S. S. Jaffer, A. Hussain, M. A. Qureshi, J. Mirza, and K. K. Qureshi, "A low cost PON-FSO based fronthaul solution for 5G CRAN architecture," *Optical Fiber Technology*, vol. 63, May 2021.

[26] D. Ulloa, G. Arévalo, and R. Gaudino, "Optimal deployment of next-generation PON for high and ultra-high bandwidth demand scenarios in large urban areas," in *22nd International Conference on Transparent Optical Networks (ICTON)*, 2020, pp. 1–6.

[27] F. Yaghoubi, M. Mahloo, L. Wosinska, *et al.*, "A techno-economic framework for 5G transport networks," *IEEE Wireless Communications*, vol. 25, no. 5, pp. 56–63, 2018.

[28] L. Chang, Z. Zhang, P. Li, *et al.*, "6G-enabled edge AI for metaverse: Challenges, methods, and future research directions," *Journal of Communications and Information Networks*, vol. 7, no. 2, pp. 107–121, 2022.

[29] R. Bassoli1, F. H. Fitzek, and E. Calvanese Strinati, "Why do we need 6G?" *ITU Journal on Future and Evolving Technologies*, vol. 2, no. 9, pp. 1–31, 2021.

[30] Z. Zhang, Y. Xiao, Z. Ma, *et al.*, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, 2019.

[31] H. Tataria, M. Shafi, A. F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G wireless systems: Vision, requirements, challenges, insights, and opportunities," *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1166–1199, 2021.

[32] C. D. Alwis, A. Kalla, Q.-V. Pham, *et al.*, "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open Journal of the Communications Society*, vol. 2, pp. 836–886, 2021.

[33] O-RAN Open Xhaul Transport Working Group 9, "Xhaul packet switched architectures and solutions," Technical specification, version 3.00, Jul. 2022.

[34] L. Cominardi, L. M. Contreras, C. J. Bcrnardos, and I. Berberana, "Understanding QoS applicability in 5G transport networks," in *2018 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2018, pp. 1–5.

[35] J.-P. Vasseur, M. Pickavet, and P. Demeester, *Network Recovery: Protection and Restoration of Optical, SONET-SDH, IP, and MPLS*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2004, ISBN: 012715051X.

[36] C. N. da Silva, L. Wosinska, S. Spadaro, J. C. W. A. Costa, C. R. L. Frances, and P. Monti, "Restoration in optical cloud networks with relocation and services differentiation," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 8, no. 2, pp. 100–111, 2016.

[37] B. Mukherjee, *Optical WDM Networks*. Springer New York, NY, 2006, ISBN: 978-0-387-29055-3.

[38] P. P. Sahu, *Optical Networks and Components, Fundamentals and Advances*. Boca Raton: Taylor & Francis Group, 2020, ISBN: 9780429298417.

[39] L. Valcarenghi, R. Inkret, B. Mikac, *et al.*, "Which resilience for the optical internet? an e-Photon/ONe+ outlook," in *9th International Conference on Transparent Optical Networks*, vol. 3, 2007, pp. 142–145.

[40] M. Held, L. Wosinska, P. Nellen, and C. Mauz, "Consideration of connection availability optimization in optical networks," in *Fourth International Workshop on Design of Reliable Communication Networks, 2003. (DRCN 2003). Proceedings.*, 2003, pp. 173–180.

[41] J. Segovia, E. Calle, P. Vila, J. Marzo, and J. Tapolcai, "Topology-focused availability analysis of basic protection schemes in optical transport networks," *Journal of Optical Networking*, vol. 7, no. 4, pp. 351–364, Apr. 2008.

[42] Millimetre wave transmission ETSI industry specification group, "ETSI GS NFV-REL 003, report on models and features for end-to-end reliability," Group specification, version 1.1.2, Jul. 2016.

[43] ITU-T, "Characteristics of transport networks to support IMT-2020/5G," G-series Recommendations, 8300, May 2020.

[44] NGMN Alliance, "5G extreme requirements: End-to-end considerations," White paper, 2019, version 2.5.

[45] E. J. Oughton, K. Katsaros, F. Entezami, D. Kaleshi, and J. Crowcroft, "An open-source techno-economic assessment framework for 5G deployment," *IEEE Access*, vol. 7, pp. 155 930–155 940, 2019.

[46] H. Frank, R. S. Tessinari, Y. Zhang, *et al.*, "Resource analysis and cost modeling for end-to-end 5G mobile networks," in *International IFIP Conference on Optical Network Design and Modeling*, Springer, 2019, pp. 492–503.

[47] W. Xie, N.-T. Mao, and K. Rundberget, "Cost comparisons of backhaul transport technologies for 5G fixed wireless access," in *IEEE 5G World Forum (5GWF)*, 2018, pp. 159–163.

[48] S. Roblot, M. Hunukumbure, N. Varsier, *et al.*, *Techno-economic analyses for vertical use cases in the 5G domain*, arXiv:1906.09746, 2019.

[49] I. Mesogiti, G. Lyberopoulos, F. Setaki, *et al.*, "Macroscopic and microscopic techno-economic analyses highlighting aspects of 5G transport network deployments," *Photonic Network Communications*, vol. 40, no. 3, pp. 256–268, 2020.

[50] ITU-T, "5G wireless fronthaul requirements in a passive optical network context," G-series Recommendations, Supplement 66, Sep. 2020.

[51] J. S. Wey and J. Zhang, "Passive optical networks for 5G transport: Technology and standards," *Journal of Lightwave Technology*, vol. 37, no. 12, pp. 2830–2837, 2019.

[52] M. H. Keshavarz, M. Hadi, M. Lashgari, M. R. Pakravan, and P. Monti, "Optimal QoS-aware allocation of virtual network resources to mixed mobile-optical network slices," in *IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–06.

[53] 3GPP, "TS 38.211, physical channels and modulation," Technical specification, version 17.2.0, Jun. 2022.

[54] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1184–1189.

[55] J. Mocerino, "5G backhaul/fronthaul opportunities and challenges," Fujitsu Technical paper, Oct. 2019.

[56] ITU-R, "Availability objectives for real digital radio-relay links forming part of a high-grade circuit within an integrated services digital network," F-series Recommendations, F.695, Jun. 1990.

[57] H. Lehpamer, *Transmission systems design handbook for wireless networks* (Artech House mobile communications series). Artech House, 2002, ISBN: 9781580535540.

[58] H. Long, M. Ye, G. Mirsky, A. D'Alessandro, and H. Shah, "Ethernet traffic parameters with availability information," RFC 8625, Aug. 2019.

[59] 3GPP, "TS 22.104, service requirements for cyber-physical control applications in vertical domains," Technical specification, version 18.3.0, Dec. 2021.

[60] B. M. Khorsandi, C. Raffaelli, M. Fiorani, L. Wosinska, and P. Monti, "Survivable BBU hotel placement in a C-RAN with an optical WDM transport," in *13th International Design of Reliable Communication Networks (DRCN) Conference*, Mar. 2017, pp. 1–6.

[61] M. Shehata, F. Musumeci, and M. Tornatore, "Resilient BBU placement in 5G C-RAN over optical aggregation networks," *Photonic Network Communications*, vol. 37, no. 3, pp. 388–398, Jun. 2019.

[62] M. Y. Lyazidi, L. Giupponi, J. Mangues-Bafalluy, N. Aitsaadi, and R. Langar, "A novel optimization framework for C-RAN BBU selection based on resiliency and price," in *IEEE 86th Vehicular Technology Conference (VTC-Fall)*, Sep. 2017, pp. 1–6.

[63] M. Klinkowski and M. Jaworski, "Dedicated path protection with wavelength aggregation in 5G packet-optical Xhaul access networks," *Journal of Lightwave Technology*, vol. 41, no. 6, pp. 1591–1602, 2023.

[64] A. F. Beldachi, M. Anastasopoulos, A. Manolopoulos, A. Tzanakaki, R. Nejabati, and D. Simeondou, "Resilient cloud-RANs adopting network coding," in *Optical Network Design and Modeling*, A. Tzanakaki, M. Varvarigos, R. Muñoz, *et al.*, Eds., Cham: Springer International Publishing, 2020, pp. 349–361.

[65] A. Alabbasi, X. Wang, and C. Cavdar, "Optimal processing allocation to minimize energy and bandwidth consumption in Hybrid-CRAN," *IEEE Transactions on Green Communications and Networking*, vol. 2, no. 2, pp. 545–555, Jun. 2018.

[66] C. I, Y. Yuan, J. Huang, S. Ma, C. Cui, and R. Duan, "Rethink fronthaul for soft RAN," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 82–88, Sep. 2015.

[67] T. Doukoglou, V. Gezerlis, K. Trichias, *et al.*, "Vertical industries requirements analysis & targeted KPIs for advanced 5G trials," in *European Conference on Networks and Communications (EuCNC)*, 2019, pp. 95–100.

[68] R. F. Vieira, P. H. Alves Pereira, and D. L. Cardoso, "Resource allocation optimization for hierarchical cloud data centers," in *4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*, 2018, pp. 1–6.

[69] M. Mosahebfard, J. Vardakas, K. Ramantas, and C. Verikoukis, "SDN/NFV-based network resource management for converged optical-wireless network architectures," in *21st International Conference on Transparent Optical Networks (ICTON)*, 2019, pp. 1–4.

[70] Y. Zhou, B. Ramamurthy, B. Guo, and S. Huang, "Resource delayed release strategy for dynamic and fast end-to-end service provisioning in SDN-enabled OTN over WDM networks," in *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, 2017, pp. 1–6.

[71] R. Mao and H. Du, "DPMA: A distributed profit-based placement scheme for multi-sp mobile edge computing," *Journal of Combinatorial Optimization*, vol. 44, no. 5, pp. 3294–3309, 2022.

[72] M. K. Singh, S. Vittal, and A. Antony Franklin, "SERENS: Self regulating network slicing in 5G for efficient resource utilization," in *IEEE 3rd 5G World Forum (5GWF)*, 2020, pp. 590–595.

[73] M. Rayani, A. Ebrahimzadeh, R. H. Glitho, and H. Elbiaze, "Ensuring profit and QoS when dynamically embedding delay-constrained ICN and IP slices for content delivery," *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 2, pp. 769–782, 2022.

[74] M. R. Raza, A. Rostami, L. Wosinska, and P. Monti, "A slice admission policy based on big data analytics for multi-tenant 5G networks," *Journal of Lightwave Technology*, vol. 37, no. 7, pp. 1690–1697, 2019.

[75] B. Skubic and I. Pappa, "Energy consumption analysis of converged networks: Node consolidation vs metro simplification," in *Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*, 2013, pp. 1–3.

[76] Y. Huang, Y. Zhu, X. Qiao, X. Su, S. Dustdar, and P. Zhang, "Toward holographic video communications: A promising AI-driven solution," *IEEE Communications Magazine*, vol. 60, no. 11, pp. 82–88, 2022.

[77] C. Wu, R. Buyya, and K. Ramamohanarao, "Cloud pricing models: Taxonomy, survey, and interdisciplinary challenges," *ACM Comput. Surv.*, vol. 52, no. 6, Oct. 2019, ISSN: 0360-0300.

[78] *Intel: 90% of 5G data will be video, but AR gaming and VR will grow*, `https://venturebeat.com/games/intel-90-of-5g-data-will-be-video-but-ar-gaming-and-vr-will-grow/`, Accessed on: 2022-12-26.

[79] M. Quagliotti, A. Rafel, O. Gonzales De Dios, V. López, *et al.*, "Definition of use cases, service requirements and KPIs," Metro-Hual Deliverable D2.1, 2018.