POLITECNICO DI TORINO

Master's Degree in COMMUNICATIONS AND COMPUTER NETWORKS ENGINEERING

Master's Degree Thesis

Machine learning techniques to reduce quality of transmission uncertainties in transparent optical networks

Supervisors

Prof. Vittorio CURRI

M.Sc. Andrea D'AMICO

Candidate

Mirna ALMOKDAD

July 2021

Summary

In this thesis we consider the problem of the uncertainty on the Generalized Signal to Noise Ratio (GSNR) value over optical networks. Optical networks are considered in a partially disaggregated approach, where each optical line systems (OLSs) separately contributes on the total capacity. Our project is based on a Python implementation where we study the GSNR distribution in realistic scenarios applying the Monte Carlo method and machine learning (ML) techniques.

In general, as in a real scenario, the connector input loss and the ripples are not characterized and cannot be exactly estimated. We study the variation of the GSNR distribution due to uncertainties over the considered parameters assuming Gaussian fluctuations around reference values for both input connector loss and ripples.

Our network topology is constructed by reconfigurable optical add-drop multiplexers (ROADMs), on wavelength-division multiplexing (WDM) C-BAND, connected by independent OLSs. Additionally, we consider each OLS composed of Standard Single Mode Fiber (SSMF) and Erbium-Doped Fiber Amplifier (EDFA). These OLSs transport lightpaths WDM C-BAND optical signals are considered as a fully loaded system from ROADM to ROADM independently. Every channel propagated along the OLS is considered separately with respect to the other channels propagated. Additionally, we focus our study on a channel under test (CUT). As we do not consider the Raman effect, we set as CUT the central channel of the LP that represents the worst case scenario by means of the nonlinear degradation introduced within the propagation through the entire OLS.

Each lightpath (LP) has been abstracted as an additive white Gaussian noise (AWGN) channel. Along the OLS at the optimal working point, the quality of transmission (QoT) of the line can be softwarized and predicted in order to optimize the transmission. Moreover, the GSNR can be used as QoT metric over each LP. The GSNR takes into account the accumulation of two main noise components: the amplified spontaneous emission (ASE) and the nonlinear interference (NLI) noises. The ASE noise is generated by the optical amplifiers used in-line as repeaters and at the receiver side as pre-amplifiers, while the NLI noise is introduced by the propagation through the fiber spans along the OLS and is always has been

calculated at full spectral load in order to avoid load-dependent network behaviors.

This enables a deeper insight on launch power dependency of the GSNR, in which the power channel is optimized for the maximum GSNR along the OLS.

To study the GSNR distributions, we analyse separately the ASE and NLI noises introduced during the signal propagation and that affect the LP.

After a comprehensive description of the statistical features of the GSNR distribution over a periodic OLS, we extend the results over an entire network. In this wider scenario, we investigate the total GSNR inaccuracy over generic lightpaths due to physical parameters uncertainty.

We create a dataset of GSNR variations using an abstraction of the OLSs that simulates the signal propagation over a real optical network topology. The generated dataset is formed by connections randomly chosen and deployed following the algorithm of the shorted path from the source to the destination ROADMs.

On top of the collected GSNR for each lightpath propagated, we prove that this quantity always have a Gaussian distribution. Therefore, the mean μ and the standard deviation σ are essential metrics for all the QoT analysis that fully characterize the GSNR statistics.

As a matter of fact, despite an accurate GSNR prediction can be produced by various quality of transmission estimators (QoT-Es), the latter requires highly precise knowledge of the physical parameter. As this condition is not guaranteed in real case scenarios, the study of the GSNR fluctuations is a valuable tool in fixing the proper system margins.

In general, if we work with a huge topology with a big number of available connections, the creation of a dataset of the GSNR on all the full end-to-end lines is very expensive in terms of computational complexity. Thus, we apply statistical analysis to prove the relationship between the GSNR on a full path with an analytical accumulation of the distribution of each node-to-node sub-path. By means of this methodology, for a set of LPs we compare the estimated μ_{GSNR} and σ_{GSNR} with the overall values using the Z-test method. We prove that the simulated GSNR distribution on the entire LP done by Monte Carlo method is completely consistent with the predicted GSNR distribution accumulated line-by-line.

The essential motivation of our study is to manage the uncertainty introduced in the GSNR computation for each LP in order to be able to enable a reliable path computation and, then, a minimal margin LP deployment.

In order to obtain a reliable LP GSNR, the nominal working points must be properly accurate. Therefore, to deploy more traffic with an optimum exploitation of the installed equipment, a limited system margin is required. Nevertheless, an adequate margin is required to ensure the reliability of the predictions to guarantee the proper system behavior in term of functionalities.

Given the GSNR Gaussian distribution, the applied margin has been chosen following the practical engineering rule of "$3\sigma_{GSNR}$" for which we expect a conservative

GSNR estimation almost on the 99% of the connections.

Beside the statistical approach, this investigated framework represents an ideal scenario for the application of ML on top of the statistics of GSNR fluctuations.

This may be achieved by collecting a dataset of the OLS responses to various spectral loads in order to train a ML algorithm, allowing a QoT-E to be calculated for both untested spectral load configurations and LPs which have not yet been explored. In particular, we apply a transfer learning by using the deep neural network (DNN) algorithm feeding it with a training dataset (based on the German topology) aiming to correct the QoT-E in an unknown network topology (the US network topology) that we consider as a testing dataset. Then we collect the predicted GSNR by the ML algorithm.

In conclusion, we compare the QoT prediction results of the statistical approach and the transfer learning on US network. In order to compare the predicted GSNRs of both the ML and the statistical approach by means of accuracy and reliability. The accuracy has been quantified in terms of the root mean squared error (RMSE) by calculating the error distance between the actual and the predicted GSNR. On the other hand, the reliability has been evaluated by calculating the percentage of conservative GSNR predictions over the total number of investigated connections, namely, the percentage of predicted GSNRs lower than the actual values.

The statistical approach is more accurate and gives a lower RMSE with respect to the ML method. Also, the reliability of the statistical method is higher with a high amount of reliable connection cases, instead the ML predicted results produce higher probability of going out of service.

Indeed, the performance of the applied ML algorithm can be increased with a more fine tuning of the hyper parameters. This improvement has not been considered in this work and would require a further analysis of the investigated scenario, especially to avoid any overfitting.

KEYWORDS: optical networks, Monte Carlo, quality-of-transmission, deep neural networks, machine learning

Acknowledgements

I want to start by praying and thanking God for keeping helping me and making my life brilliant, looking forward for achieving my milestones.

I am deeply grateful to my tutors, Professor Vittorio Curri and Andrea D'amico. With their precious advises, their scrupulous and punctual support, their in-depth knowledge on the subject and availability, helped me in every step of my thesis.

Therefore, I dedicate this thesis to MCF Charity. In specific, I thank from my heart Dr. Nadia and Mrs. Dania Cheaib that helped me to reach my dream.

I warmly thank my mom Salwa and my sister Zinat that think of me in their prayers every moment and scarified their life to serve my dreams. Without them I would not be here now!

Thanks to my relatives and my second family Salma, Kassem and Leila that followed my lacks and needs and hugged me as their daughter always. With their love, this amazing milestone is becoming possible.

From all my heart, I say 'Grazie mille' to my best friend Matteo. With his kindness and his love, he shared with me all my moments, putting as a first priority my happiness and my success.

Moreover, I must thank Paola. She became a second mom and a friend for me in Italy. She cared and helped me to spend in a chilled way my stressful time, by her tendency; with her I felt always at home!

Finally, I must recognise the sentimental value the true friendship has in my life. Abbass, Hoda, Marwa and Zalfa you were always by my side, step by step, celebrating with me my achievements.

I see this thesis as a start of my future career and and I hope I will have a brilliant path ahead of me.

"Ad maiora semper!"

Table of Contents

List of Tables

List of Figures

Acronyms

WDM
: Wavelength division multiplexing

DWDM
: dense wavelength division multiplexing

OEO
: optical-electricaloptical

SDN
: Software defined network

ODTN
: Open and Disaggregated Transport Networks

ROADM
: reconfigurable optical add-drop multiplexer

OLS
: Optical line system

LP
: Light path

API
: application programming interface

EDFA
: Erbium-Doped Fiber Amplifier

OLC

OLS controller

GSNR

Generalized signal to noise ratio

QoT

Quality of transmission

MSA

MultiSource Agreement

QoS

quality of service

BER

bit error rate

DSP

digital signal processing

OTN

optical transport network

PHY

physical layer

R_b

Bit rate

R_s

symbol rate

PLI

Physical linear impairments

SPM

self-phase modulation

XPM
: cross-phase modulation

FWM
: four-wave mixing

ASE
: Amplifier Spontaneous Emission

PNLI
: physical Non-linear impairments

NF
: noise figure

ISI
: inter-symbol interference

GN
: Gaussian-Noise

NLI
: Non Linear Interference

RI
: Rand Index

SC
: Size of Class

SPN
: Sensor Pattern Noise

SSC
: Sparse Subspace Clustering

AWGN
: additive White Gaussian Noise

CLT
: Central Limit Theorem

DeT
: Deutsche Telekom

L-PCE
: lightpath computation engine

ONC
: optical network controller

RWA
: Routing and wavelength algorithm

TAPI
: Transport Application Programming Interface

DT
: Digital twin

RMSE
: Root Mean Square Error

NE
: Network element

PDF
: probability density function

DNN
: deep neural network

Chapter 1

Introduction

In general, data transmission was develop based on two fundamental types radio and wire signals. This data transmission model has changed by the invention of the fiber optics from the 1980 up to now.

In the world of open optical network domain, the communication system is based on sending signals as lightpath in order to transfer information from a give source to destination address. This communication can scale according to the special need, starting from small connection between computers in offices going to cover the world. The demand on the internet traffic is hugely increasing every year in a dynamic way, and it requires a flexible response in term of quality of transmission (QoT) and of service (QoS) and network efficiency performance.

Nowadays, the optical transmission is developed in a new paradigm introducing new protocols as the WDM that ensure the transferring in an efficient way along the optical network. The researchers proved that a network will work without human interaction. That was possible by the introduction of SDN facilities that enhance the flexibility in the network wavelength allocation and provide a huge bandwidth satisfying the traffic network demands and enhance the QoS.

In fact, a very important role in optical transmissions goes to the amplifiers included along optical line system (OLS) with the fiber spans. Thanks to their ability of restoring the losses, they restore the optical input power even after propagating a lightpath over a very long routing path.
Starting from a general overview of the optical domain, we describe the physical layer abstraction for our network topology introducing the optical fiber links.

Our work presents a way of describing methodologies and algorithms based on optimization tools for the network parameters. We discuss real network topologies under examination with the relative analysis. Accordingly, results are obtained during the research activity showing that the prediction of the network performance by applying a margin of conservation satisfies the needs of the real network simulations with impressive performances.

The extraordinary growth in the network traffic in the optical communication domains implies the introduction of Machine learning techniques (ML). The ML is able usually to learn from a topology, by a specific method, then predict the QoT-E estimation and the performance of an unestablished propagated lightpath.

ML is a state-of-the-art technique that is used to analyze the traffic along the network and chooses the useful data from it and makes decisions. Also, by estimating of (QoT) is essential for lowering supplied margins and for optimizing the optical network design.

1.1 Contributions and organization of the thesis

In this thesis we consider the problem of variation of the GSNR distribution due to uncertainties over many parameters assuming Gaussian fluctuations around reference values. The thesis is organized in the following chapters:

- Chapter 2: **Background on fiber optics** presents an overview of open and disaggregated optical networks, presenting the WDM modulation, the concept of Software Defined Networks and the Digital Signal Processing algorithm. In this section, we also present the existing state of the art works in the open optical domain.
- Chapter 3: **Physical layer abstraction** describes the data plane managed by the OLS controller. The main focus is about the physical linear impairments as: Chromatic Dispersion β_2, Amplifier Spontaneous Emission Noise and the power loss. While, the non-linear impairments that have the higher effect on the channel degradation done by the nonlinear crosstalk among channels and it is induced by the Kerr effect. The NLI is an additive Gaussian random process that adds up to the ASE noise contributing to the overall GSNR degradation as a unique parameter of QoT.
- Chapter 4: **The activities** describe our meshed weighted topology. It shows an example of a real scenario of Telekom in German, in which the measurements are done by having an overview of the logical connection between the cities and their positions. Then creating 2 dataset: dataset1 that contain all the information related to full path connections, instead dataset0 contains the

node-to-node connections description.

The activities is split between sections in which, in **section 1** we describe our system operating system, that is partially aggregated under the orchestration of an SDN paradigm.

In **section 2** , we present the Montecarlo model applied to study the QoT for our network topology.

Further, **section 3** describes our network architecture as physical layer abstraction, and the positioning of our network elements along the optical line systems.

Section 4 represents the experimental study to calculate the GSNR for each dedicated propagated lightpath in the routing space.

Instead, in **section 5** we present the analytical method that compute the predicted GSNR by regression along the sub-paths of a full OLS.

Then, in **section 6** we prove graphically and statistically that our generated GSNR distribution ,done by Montecarlo runs for every connection, is Gaussian. The statistical normality test are : "Shapiro-Wilk Test" and "D'Agostino's K^2 Test".

Moreover, **section 7** applies the "Z-test" in order to check the similarity in the normal distribution done analytically and experimentally for the generated GSNRs.

Also, **section 8** describes the applied conservative margin to put an optimal point for the analytical generated GSNR. The margin applied by the "3-sigma limits". Then after applying the margin we count the percentage of going out of service in our system.

Finally, **section 9** studies the difference between the optimal and the applied GSNR for every lightpath propagated by computing the corresponding RMSE for all the generated connections.

section 10 Transfer learning application from a known network topology to an unknown topology to obtain a QoT-E without uncertainties. Then application of statistical analysis for the prediction of GSNR with some margin of conservation to adjust the ML results.

- ☐ Chapter 5: **Results** contains a discussion about the experimental and analytical studies for our optical network topologies and the evaluation metrics. It also analyzes the results in tables of performances and graphs. The final section compares the proposed methods among themselves and with a way of optimizing the QoT to create a margin of conservation for the network performance with an application of ML.
- ☐ Chapter 6: **Conclusions** concludes the work with a comment about the results obtained and some possible future refinement work to be implemented over our activities of predicting and optimization.

Chapter 2

Background on fiber optical networks

2.1 History of communication for open and disaggregated optical networks:

Optical fibers represent a fundamental technological medium for modern communications, thanks to the development stage at which optical fiber links currently are. Optical fibers have been already deployed over the continents and across the oceans and they guarantee fast access between very far receivers. The optical-fiber applications are numerous; they are diffused in offices and homes, so they can really help to relate people and servers everywhere on Earth.
These applications involve the transmission of data, video or voice over distances of less than a meter to hundreds of kilometers, using just few standard fiber designs and several cable designs.

The request of data transferring and the capacity on the traffic demand along the optical networks in an efficient manner nowadays grown exponentially and opens new challenges. We require an answer to this capacity increasing demand, we have to rely more on fiber optics than on copper cables that are still extensively used for data transmission.
Multinational firms need secure, reliable systems to transfer data and financial information between buildings to the desktop terminals and to transfer data around the world.

The very high bandwidth provided by optical fibers makes it a subject of perfect use for transmitting broadband signals, such as high-definition television (HDTV)

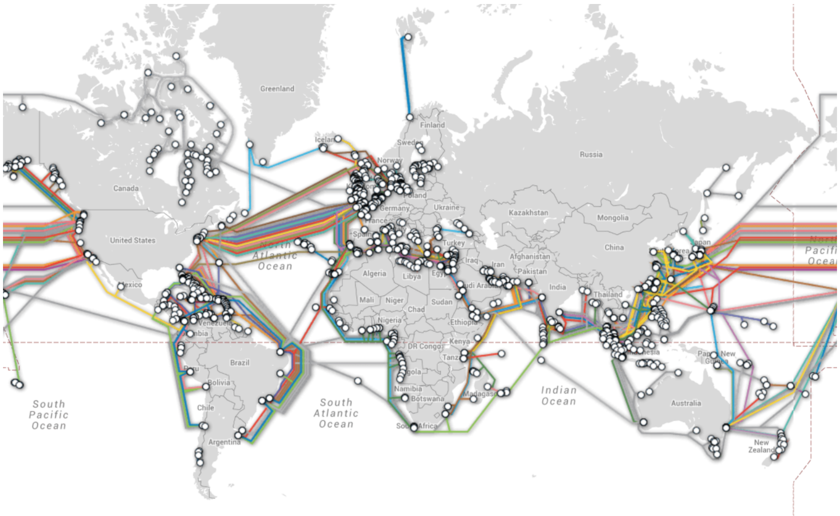

Figure 2.1: Optical Fibers distribution over the globe

telecasts [1].

Optical fibers are used by telecommunications companies to transmit telephone signals, Internet communication and cable television signals.
The main advantages of optical networks are that they have high speed capability, they can send up to 50 Terabits per second using a single fiber. Moreover, they require low power and they have low signal attenuation and low signal distortion.

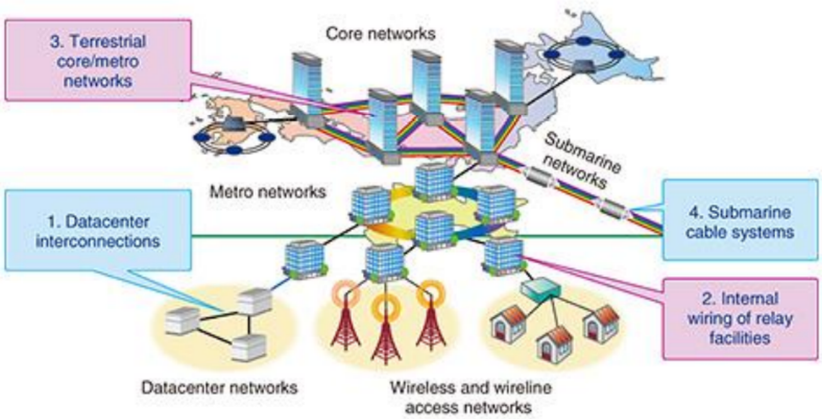

Figure 2.2: Optical Fibers design

An optical fiber is a very thin cylinder made usually in glass since it has very low attenuation values at optical frequencies, consisting of a central “core” surrounded by a “cladding” in which the core has a refractive index slightly higher than the cladding as seen in figure 2.3 [2].

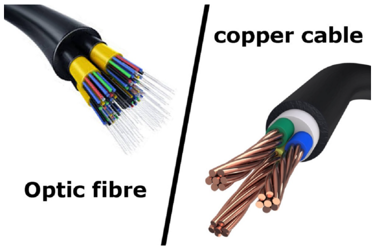

Figure 2.3: Comparison between a optical fiber and a copper cable

An optical fiber is single-mode if its Normalized Frequency Parameter ν is below a threshold value $\nu \leq 2.405$.
Single-mode fibers are used for transmission distances longer than 1 km that’s why in our work we focus on single mode fibers for which the multi-mode fibers are used for LAN networks [2].

This optical communication system uses the light signals in order to transfer data between nodes which is done over a network called "Switching nodes". These nodes may be connected to nodes or to stations.

The switching nodes concern on routing the data from node to another, up to when it reaches the destination.
The technique used in this communication system is the packet switching because of the line efficiency, data rate conversion and the flexibility in which even heavy packets are accepted avoiding blocking events.
After 2010, optical networks changed to a new paradigm in which they become Elastic Optical Networks EON [3].

2.2 wavelength division multiplexing modulation technique

The wavelength division multiplexing (WDM) is a main modulation system used in order to maximize the capacity of the transmission in an optical fiber transmission system. In such system, a single optical fiber may be used to carry multiple optical signals in what is called a wavelength division multiplex system which is which is a very well-known approach.

As nowadays, the future of the networks channels is reaching a wide-area of a backbone, the employment of the dense wavelength division multiplexing (DWDM) and wavelength routing in optical networks is becoming an essential need to be ready as a candidate for that future [4].

This employment of DWDM introduces the concept of reconfigurable wavelength add/drop in nodes that are referred to as optical add/drop multiplexers or (ROADM).
With this feature of wavelength add/drop function, particular wavelengths could now be reconfigured, added and dropped in pre-planned network configurations.
A degree represents a direction in which the node is able to connect to another

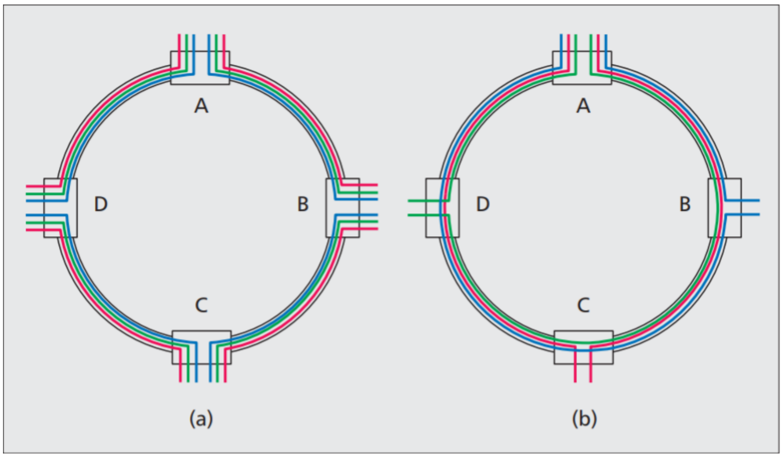

Figure 2.4: Basic DWDM functions: a) multiplexing/demultiplexing; b) add/drop

node. Figure 2.4 shows a basic architecture of a ROADM having four degrees.2.4
In order to carry in a single medium multiple data signals, the multiplexing is applied on these data signals by the WDM modulation technique.
This technique is done by dividing the available bandwidth of transmission on a fiber optic into many different smaller channels treated as capacity bandwidth

non-overlapping. For that reason, each of these channels is assigned to a different carrier wavelength λ as shown in figure 2.5.

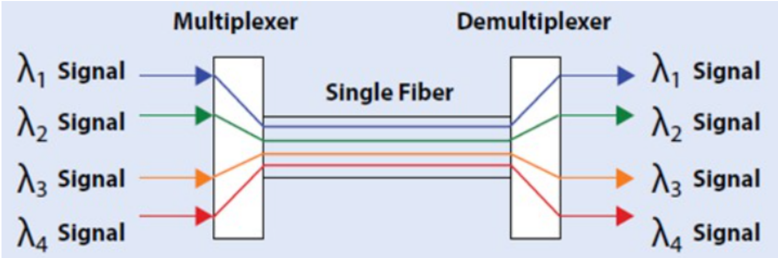

Figure 2.5: Wavelength division multiplexing scheme

In fact, these lightpaths are transparently routed in optical domain at a given wavelength over the network as paths, and it is shown in figure 2.6 different wavelengths are multiplexed within a fiber sharing the same medium single fiber, then demultiplexed as the original deployed lightpath. In general, these routes are defined by the switching matrices of switching nodes for each lightpath apart over the passing by the inner hops as optical switching as shown in figure below 2.6 .

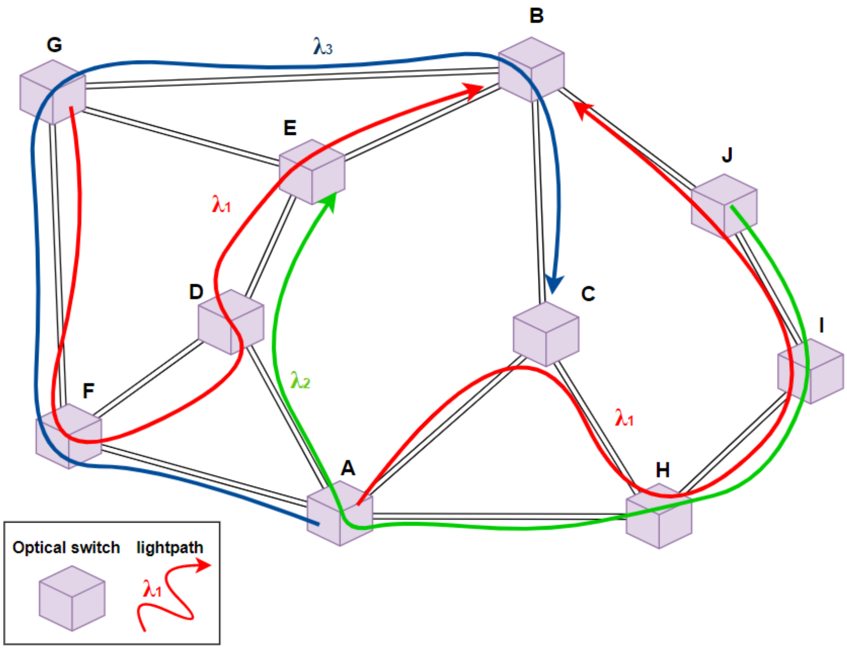

Figure 2.6: Lightpath propagation over a network topology

All these dedicated lightpaths are on demand resources assigned as channels or circuits from source to destination for every involved transceiver. As soon as this lightpath reaches the desired destination, and the communication finishes, the network turns down this assigned lightpath [2, 5].

Moreover, when a lightpath is established between any two nodes in a transparent optical network, the reserved traffic between these nodes can be routed without buffering and it does not require any intermediate optical/electrical-optical (OEO) conversion [4].

2.3 Software defined network:

This transparent optical networks evolution moved toward the implementation of the openness paradigm; an architecture to control not just a networking device but an entire network, which is called software defined network (SDN).

The SDN architecture constructed with three layers:

1) Infrastructure layer (Data plane layer)

2) SDN control layer (control plane layer)

3) Management plane

In principle, SDN separates and decouple the control plane and data plane entities and it manages the routing and network configuration.

The controller becomes a brain orchestrator that controls and program the data plane in a centralized approach which implies that the control plane must have a global view of the optical network information. It manages and assigns rules for the data plane and is responsible for managing resource connections in the network.

Moreover, the data plane (also called transport plane) contains a set of optical or other types of switches that perform switching data in the network by processing and delivering the packets in a local forwarding state. The control plane has to decide the routing algorithm and the destination of these packets by centralized computations using the OpenFlow protocol.
These switches are connected by physical links in the optical network topology. This plane will receive orders from the control plane, using the rules implemented in forwarding tables.

The management of the control plane is done by the Management plane. Also, it is responsible for configuring the routing area, the control plane resources and transporting data in the control plane.

The main reason of the SDN is because networks are hard to manage, so it creates a more flexible and manageable infrastructure by virtualizing the computation and the storage.
Moreover, the advantages involves also the easier operation over the network with the increase of the scalability and the reliability in which the design and the planning is way simpler.

2.4 Optical Line System Controller:

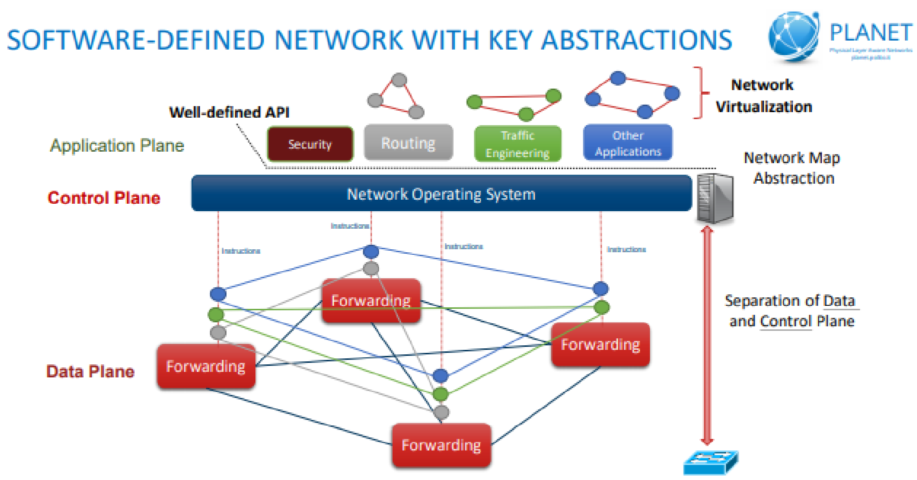

Figure 2.7: SDN separation of control and data planes

Within the SDN architecture, an OLS controller (OLC) sets the working model of each lightpath assigned to an OLS as the gain of each booster preamp amplifier of the line, the degree unit of the source and destination ROADMs along with the in line amplifiers as shown in figure 2.7 the decoupling in SDN between control and data plane.

Moreover, it sets consequently the power per channel at the input of each fiber span and to optimize transmission (maximum Generalized signal to noise ration (GSNR)).
In fact, The Quality of transmission (QoT) of the OLS line can be then predicted and softwarized.

2.5 Fully disaggregated optical networks for Open and Disaggregated Transport Networks:

The application of the openness paradigm SDN needs the optical network to be partly or fully disaggregated. A fully disaggregated network means that each network element is accessible by the SDN controller with common protocols, data structures and control APIs [6].

In such a fully disaggregated model, the transport system is also disaggregated, and the optical network elements ROADMs can be provided by different vendors with open and standard APIs directly to the optical SDN controller. These APIs can be based on the Open ROADM MultiSource Agreement (MSA).

The disaggregation goal is to provide a degree of flexibility such as component migration. Open and Disaggregated Transport Networks shorten by (ODTN) are implemented by DWDM systems, including transponders, Open Line Systems, amplifiers, multiplexers, all-optical switches and ROADMs.
The combination between the disaggregation and software defined networks provide a key point to the simplification and the automation of network operations.
SDN facilities underline and enrich the flexibility in wavelength allocation and increase the backbone bandwidth to satisfy the quality of the service (QoS) for connections and the QoT requirements for that transmission [7].

2.6 Digital Signal Processing DSP:

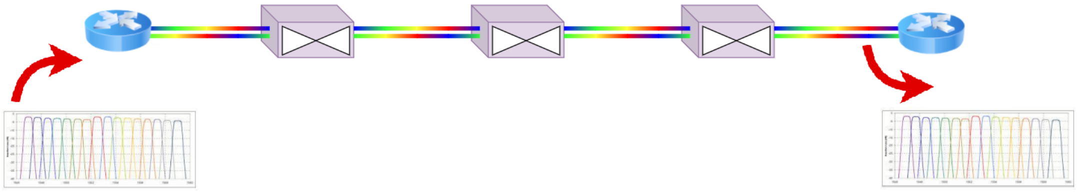

Figure 2.8: Transparency in OLS propagation

The digital signal processing (DSP) is implemented in order to compensate to the loss of power from the transmitter to the receiver ROADM by equalization of the power.

The DSP based receiver compensates and recover the linear impairments such as the chromatic dispersion apart from the electrical noises and the ASE noise. Transceivers based on coherent technology compensate for all linear impairments in fiber propagation and so as consequence the bit error rate (BER) only depends on the amount of noise introduced by the fiber propagation.

When the network is completely transparent, the routing of the wavelength in which any-to-any optical transmission is possible is done without any optical-to-electrical conversion. Moreover, the novel paradigm becomes optical transparent from the transmitter to the receiver.

The transparency in an OLS propagation introduces the fact that the optical spectrum at the output of an OLS is equal to the input spectrum plus some amount of noises as shown in figure 2.8.

The introduction of the latency in OLSs is done in optical transport and can be abstracted by its noise impairments [2] .

Chapter 3

Network physical layer abstraction

The physical layer (PHY) of our setup is the optical transport network (OTN) and it is implemented as optical transmission. The control and abstraction of the PHY layer is used to optimize the network management.
The physical layer is dependent on the network type, reach access and the type of the network as shown in the table below [4].

Network Type	Reach access	Type of element
opaque	metro	fiber spans
translucent	core haul	amplifiers
transparent	long haul	ROADM switches

Table 3.1: Network fundamental elements

Hence, there exist always techniques that are able to provide information of the PHY layer to the control plane in an SDN network. This control plane will run protocols that exploits this information efficiently in order to compute feasible routes and wavelength according to routing algorithm as mentioned before in section 3.2.

3.1 OLS controller:

The introduction of SDN disaggregated control for the physical layer PHY in a weighted graph have been included the need of an abstraction of the network infrastructure according to the QoT degradation that translated the corresponding weighted graph.

The OLS controller OLC is an essential part of the network control plane, and sets the working point of each amplifier, and consequently the input power of each fiber span along the OLS.

The WDM links over an OLS may introduce impairments into the signal path, consequently the determination of the maximum transparency length in terms of maximum distance or number of hops that an optical signal can travel is completely detected by a receiver without requiring any conversion OEO. This transparency reduces the possibility of the interaction at intermediate switching nodes along the path between the electrical layer with the optical layer.

Hence, any distance traversed by a lightpath on a specific channel over an optical path OLS depends on the following parameters:

a) the input optical signal power assigned as power per channel.

b) the length of each fiber span along the OLS.

c) the type dedicated for each fiber that can be: SSMF, LEAF or TrueWave.

d) the number of wavelengths on a single fiber for a WDM module.

e) the bit-rate (Rb) per each assigned wavelength for a lightpath.

f) the applied amplification technique and the number of amplifiers added for the compensation.

g) the type and number of switching elements (nodes as ROADM) through which the signals pass along the OLS before reaching the egress node or before regeneration.

h) The design of links involving the chromatic dispersion compensation and the loss coefficient [4].

The introduction of optical transparency in the physical layer impairment (PLI) leads to obtain a flexible and dynamic optical layer having the possibility to include intelligence such as fault management and optical performance monitoring. Hence,

some PLIs are only for transparent networks such as crosstalk impairment.
It surely has impacts on the network design by adapting the size of the transparent domains to WDM system so there is no impact of PLIs in the design process.

There exist two categories of physical impairments:

a) physical Linear impairments (PLI).

b) physical Non-linear impairments (PNLI).

The linear impairments are static and they don't depend on the intensity in fiber optics [8]. While, a non-linear impairments means intensity-dependent and it's dynamic in nature.
The main non-linear impairments that happens along an OLS and we will focus on is the Non Linear Interference (NLI) which is mainly produced by the Self-Phase Modulation (SPM) and the Cross-Phase Modulation (XPM). Instead, the linear impairments that we consider in our work are the power losses, the Amplifier Spontaneous Emission Noise (ASE) and the chromatic dispersion β_2 [8].

3.1.1 Linear impairments:

In nature, the noise is a very important and difficult topic to describe and quantify. Noises can be described as a random deviation of a physical parameter from an expected value.
In communication systems, where electrical radio, or optical signals are transmitted, the noise is considered as an impairment consequenced as a degradation and an abasement of the signal information.

In an ideal case, where communication channels are noiseless, the achievement of noiseless amplifiers and detectors becomes possible.
Hence, we will be able to communicate over wild and very long distances with very small amounts of power and with a quite null latency. In such case, we are sure there is no limit to the amplification that can be done on a signal.

As already mentioned before the noise is described as physical impairments that is represented and defined by the linear and non-linear impairment, having in mind that a transparent lightpath IS NOT practically impaired by linear propagation effects that are completely compensated for by the DSP[2].

Chromatic Dispersion:

The chromatic dispersion is one of the main kerr-effects that causes degradation on the spectral components of an optical signal during the propagation over an OLS.

This degradation is a very critical linear impairment dispersion for systems with a bit rate higher than $R_b = 2.5\,Gbps$ and so it's obvious that this dispersion depends on the fiber parameters as the material properties, bit rate, the type of the fiber and the modulation format used. These contributions approximately add up to produce the whole dispersion [4].

The chromatic dispersion β_2 adds a distortion mainly caused by the mismatch between the spectral components. The dispersion β_2 can be expressed as distortion D which is related to β_2 by a derivation of the wavelength and multiplying by dispersion:

$$\beta_2 \, [\frac{s^2}{m} \times 10^-27] \tag{3.1}$$

$$D = \frac{dW}{d\lambda}\beta_2 \, [\frac{ps}{nm\,km}] \tag{3.2}$$

The whole dispersion end to end on an OLS of a lightpath is the summation of the dispersion on each fiber link assigned to the lightpath in which this last dispersion on a fiber link is the aggregation of dispersions caused by fiber spans that frame the correspondent link.
In case of N fibers are present and dispersion is constant over each fiber what we get as total dispersion is the summation of the dispersion D_i over all the fibers over the path correspondent length L_i, as presented in the formula below:

$$D = \sum_{k=1}^{N} D_i L_i \tag{3.3}$$

Amplifier Spontaneous Emission Noise (ASE):

The optically amplified systems suffer the addition of the primary source of noises ASE which is the main contributor to the LP QoT degradation. This ASE is inserted by the optical amplifiers for which these amplifiers are used in line as repeaters and at the receiver side as preamplifiers, depending on the working points of the erbium-doped fiber amplifiers (EDFA)s within the OLSs[2, 3].

The noise figure (NF) is a parameter that mainly quantifies the noise. It is a factor that shows how much is amplified the noise power spectral density at the output end with respect to the input noise power spectral density multiplied by the amplification factor and is often specified in decibels (dB)[4].

The amplifiers emit the ASE noise in both the forward and backward directions, while only the forward ASE noise is a critical parameter taken into consideration in affecting the link performance since this noise co-propagate with the signal for which it degrades the system performance.

The generated ASE in optical amplifiers, limits the achievable gain of the amplifier and increases its noise level.
The ASE noise mixes with the optical signal and produces beat noise components at the square-law receiver[4]. AS consequence, any ASE noise introduced by each amplifier is statistically independent of the others, which means it accumulates incoherently over the line, permitting a spatially disaggregated approach.

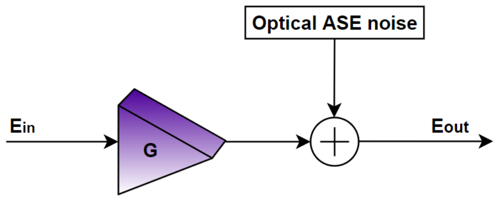

Figure 3.1: Amplifier abstraction

$$p_{signal,t} = (Flat_{loss} * Loss_{fiber}) * Gain_{amplifiers} * p_{signal,t-1} \tag{3.4}$$

$$p_{ASE} = h * f_{ref} * B_n * NF * (Gain_{amplifier} - 1)[W] \tag{3.5}$$

Power Loss:

The power loss can be specified as the optical loss that is collected from source to destination between switching nodes of the optical line systems along the fiber link in the optical networks and is normally made up of intrinsic fiber losses and extrinsic bending losses.

Intrinsic fiber losses are due to main reasons as: attenuation, absorption, reflections, refractions, Rayleigh scattering, optical component insertion losses.
Considering P_{ch} as the power launched at the input of a fiber of length L; then the output power P_{out} is given by the relation between p_{in} and α where α is the fiber attenuation loss coefficient.

The insertion loss is the loss introduced by the insertion of optical components, such as couplers, filters, multiplexers/ demultiplexers, and switches, into the optical communications system. This inserted noise is usually independent of wavelength. The extrinsic losses are due to micro and macro bending losses. In this case, additional losses occur due to the combined effects of dispersion resulting from inter-symbol interference (ISI)[4].

The power loss will be propagated along the optical line system by adding the loss generated by the connector inputs, followed by the loss introduced at the beginning of the fiber NLI and the ASE noise introduced by the amplifiers as a penalty of amplification.

In a transparency system, we assume that the loss introduced in the optical line system during the propagation is completely compensated by the gain introduced by the amplifiers along the link.

At the end of the propagation we should obtain at the output at the destination side a power more or less equal to the input power at the transmitter side regardless the perturbation and the noises happening during the propagation.

3.1.2 Non Linear impairments:

Non Linear Interference NLI:

Nowadays, the fiber attenuation and fiber dispersion are the worries that are plagued in the optical fiber communication over a propagation of a wavelength from source node to a destination node along an OLS.
These points of issues however are handled by different dispersion compensation techniques such the DSP and the addition of amplifiers over the OLSs.

However, the fiber non-linearities introduce diverse domains of drawbacks that must be solved. The consequences of non-linear impairments are becoming critical as the transmission lengths, transmission rates, number of wavelengths and optical power levels increase in addition to reduction in channel spacing [9].

In optical fiber the non-linear effects occur for two main reasons:

1) due to change in the refractive index of the medium with optical intensity (power).

2) due to inelastic-scattering phenomenon.

A general classification of non-linear effects in fiber medium is shown in figure below 3.2.

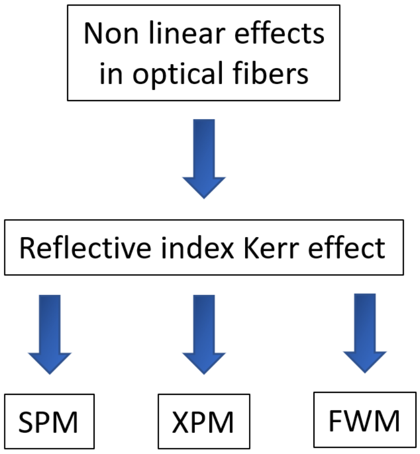

Figure 3.2: Non linear impairments

The dependence of the refractive index on the power is responsible for Kerr-effects which produces three different kinds of effects depending on the type of input signal as it's shown in figures below where Cut represents the central channel of the lightpath in which we have the worst(minimum) GSNR:

1) self-phase modulation (SPM):

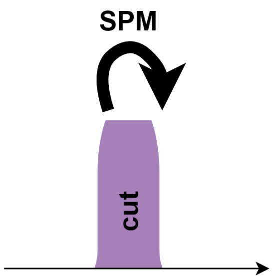

Figure 3.3: self-phase modulation

$$P_{SPM,i} = \eta_{SPM} P_i^3 \tag{3.6}$$

2) cross-phase modulation (XPM):

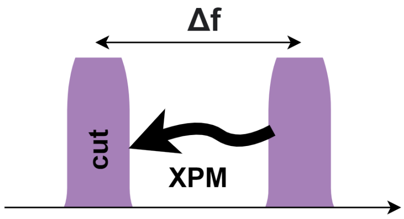

Figure 3.4: cross-phase modulation

$$P_{XPM,i} = \sum_{\substack{i=1 \\ i \neq j}}^{N_{ch}} \eta_{XPM,ij} P_i P_j^2 \tag{3.7}$$

3) four-wave mixing (FWM):

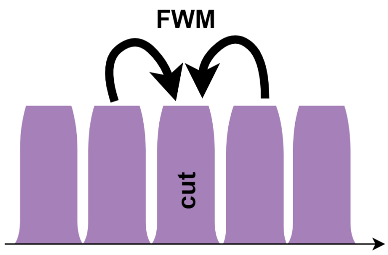

Figure 3.5: four-wave mixing

$$P_{FWM,i} = \sum_{j} \sum_{k} \sum_{l} \eta_{jkli} P_j P_k P_l \tag{3.8}$$

As consequence, the final summation of the non linear effects NLI including each of kerr-effects: SPM,XPM and FWM, is induced by the fiber propagation is in formula below: [[2]

$$P_{NLI,i} = \eta_{SPM} P_i^3 + \sum_{\substack{i=1 \\ i \neq j}}^{N_{ch}} \eta_{XPM,ij} P_i P_j^2 + \sum_{j} \sum_{k} \sum_{l} \eta_{jkli} P_j P_k P_l \tag{3.9}$$

All nonlinear effects, except SPM and XPM, provide gains to some channel at the expense of consuming power from other channels.

The importance of non-linear effects is growing due to:

1) The increase of power levels in optical domains in order to enhance the optical reach over long distances around the world.

2) The creation of more flexible networks recently by the developments in optical components such as EDFA and DWDM systems.

3) The increase in the channel bit-rate to enlarge the traffic carrying capacity of wavelengths.

4) Increase the number of wavelengths by decreasing the channel spacing to improve the overall network capacity.

3.1.3 Gaussian-Noise model assumption:

The NLI depends on the power per channel $P_c h$, on the channels active, on the channel spacing, and on the symbol rate R_s. Hence, the amount of NLI is always calculated at full spectral load in order to avoid load-dependent network behaviors.

In the Gaussian-Noise (GN) model, we have a perturbative approach which models the power spectral density of the NLI assumed to be an additive Gaussian noise disturbance with Gaussian statistics assumption for all the active channels. In the worst-case assumption, the chromatic dispersion causes over a short distance in a few kms, all modulated channels are assumed to be Gaussian modulated.

In general, over a perturbative approach the optical transparency over OLSs is also introduced. We assume to have uniform power per channel in sense that all channels are assumed to carry the same amount of power Pch and it has no frequency-dependent effect.

In such model, the NLI predicted by the GN model refers to the center channel "Cut", channel under test, supposing all the channels are active. This channel is called the cut channel and it is used for all the activities in the implementation and calculation of the parameters in the OLS to overcome the worst case scenario

of the minor GSNR.
The NLI is considered a Gaussian noise component generated by the nonlinear interaction of WDM channels.
In the analytical approximation, the NLI formula has the power spectral dependent on the number of WDM channels N_{ch}, power per channel P_{ch}, symbol rate R_s, WDM channel spacing Δ_f, and also the fiber parameters as the loss and dispersion.

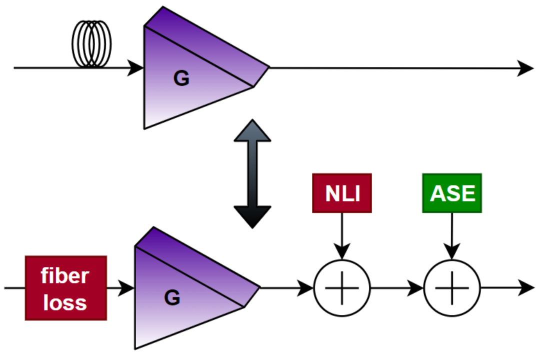

Figure 3.6: Fiber and amplifier abstraction

$$Loss = 10^{-\frac{\alpha_{dB}}{10}L_s} \tag{3.10}$$

$$L_{eff} \leq \frac{1}{2}\alpha \tag{3.11}$$

$$\eta \leq \frac{16}{27\pi} * \log\left(\frac{pi^2}{2} * \beta_2 * \frac{R_s^2}{\alpha} * N_{ch}^{\left(2*\frac{R\_s}{df}\right)}\right) * \left(\frac{\alpha}{||\beta_2||}\right) * \gamma^2 * \left(\frac{L_{eff}^2}{R_s^3}\right) \tag{3.12}$$

$$p_{NLI} = \eta * p_{signal}^3 * Loss * Flat_{loss} \tag{3.13}$$

The L effective L_eff is the length of the fiber in which we have all the effects of the non linear interference and it's related to α loss coefficient.

By looking at the analytical approximation of the NLI we clearly see that the NLI increases in cases where we have [2]:

1) The reduction of chromatic dispersion β_2 as the chromatic dispersion is inversely proportional to η.

2) Reduction of fiber loss.

3) Reduction of channel spacing.

4) Increase of nonlinear coefficient.

5) Increase of number of channel.

3.1.4 Generalized signal to noise ratio:

With the OLS at the optimal working point, the QoT of the line can be softwarized and predicted in order to optimize the transmission.
Moreover, the parameter that characterize the QoT over a lightpath is considered the generalized signal to noise ration (GSNR).
This GSNR is the same of signal to noise ratio SNR but it is more generalized, and it takes into account the accumulation of two noise components: noise ASE and NLI presented in the figure below3.6.

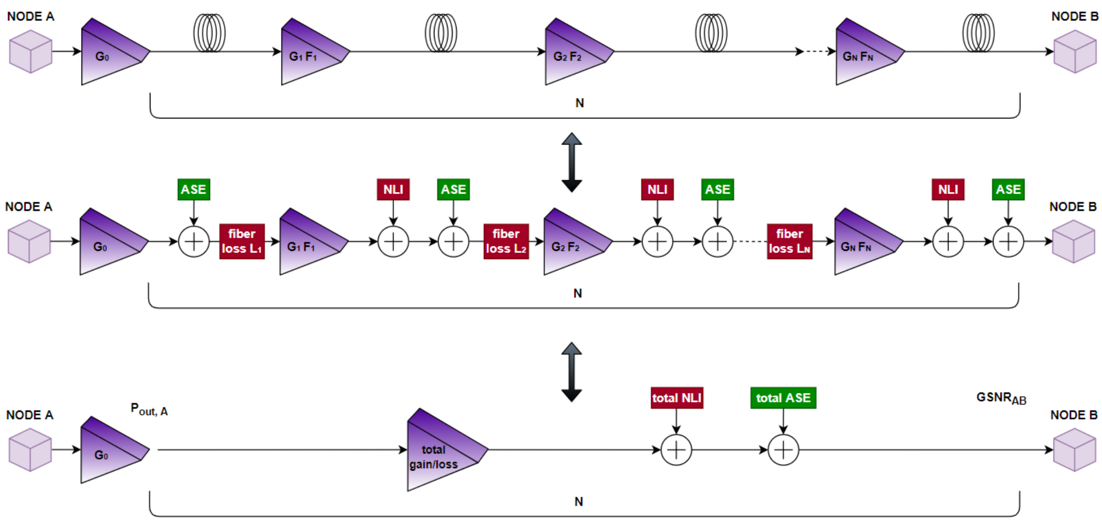

Figure 3.7: Optical Line System abstraction with ASE and NLI

$$GSNR = \frac{P_{ch}}{P_{ASE} + P_{NLI}} = \frac{P_{ch}}{P_{ASE} + \eta p_{ch}^3} \tag{3.14}$$

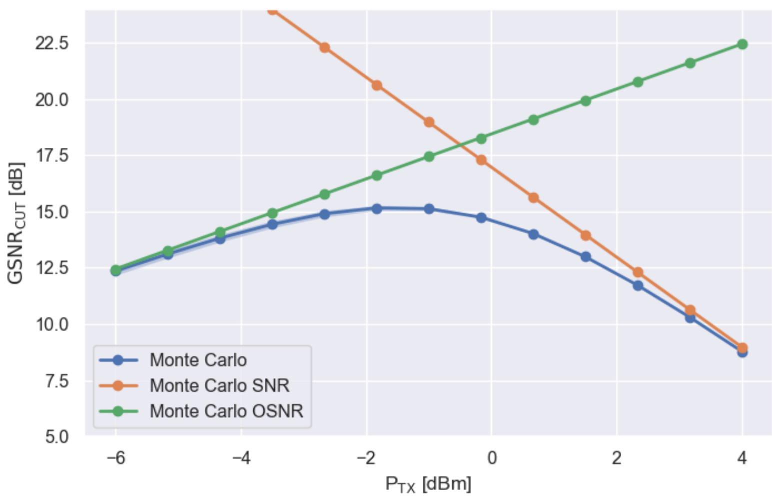

Figure 3.8: GSNR over power sweep with ASE and NLI

The GSNR can also be derived from the SNR and the OSNR as where P_{ch} is the power per channel:

$$GSNR = (OSNR^{-1} + SNR^{-1}) \tag{3.15}$$

$$OSNR = \frac{P_{ch}}{P_{ASE}} \tag{3.16}$$

$$SNR = \frac{P_{ch}}{P_{NLI}} \tag{3.17}$$

The OLS controller sets the working model of each amplifier and it aims at maximizing performance by maximizing the GSNR.
In addition, The OLS controller aims at setting the P_{ch} at the input of every fiber span to maximize the GSNR over the line and as a consequence increase the R_b. The GSNR is maximum at the optimal power that presents an optimum trade-off between the ASE and NLI noises as shown in figure 3.8.
We can see that the ASE and NLI lines are tangent to GSNR and then we obtain an optimum point because each line goes in a different direction and they cross on an optimal point. Over a general OLS, the GSNR will take into consideration all the losses, disturbance, noises and the gain introduces by the network elements. Going from an input node A to an output node B with N fiber spans, the GSNR over such OLS can be written as shown by the formula below for an end-to-end OLS[2]:

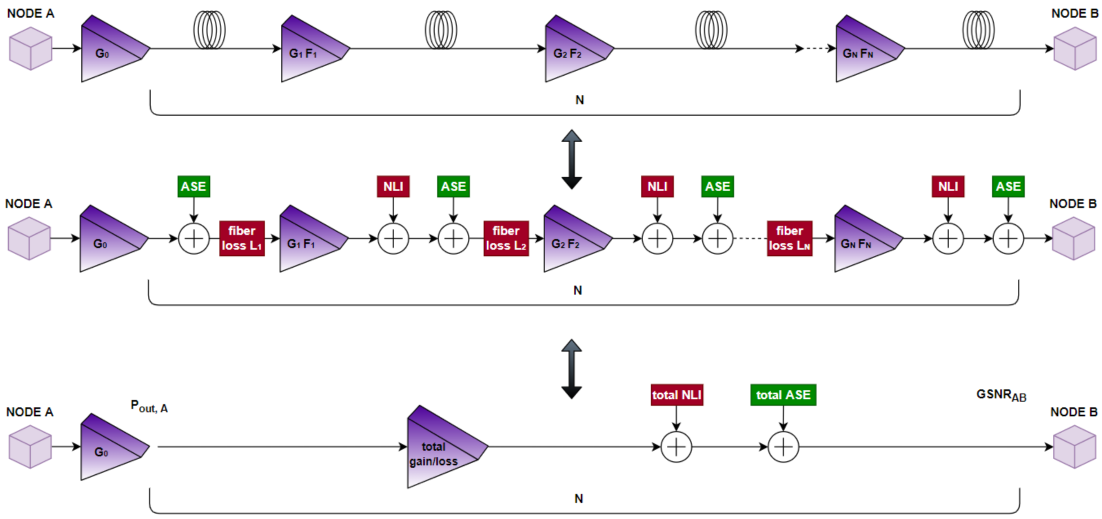

Figure 3.9: Optical Line System abstraction with ASE and NLI

3.2 The routing algorithm:

The routing algorithm is an operation to select a path from the available paths within the network to provide the best possible and available route between source and destination. The elements that should be taken into account in order to acquire this approach are:
The SDN architecture constructed with three layers:

1) The Cost must be chosen for a path in a way to not add additional cost in.
2) The distance traversed by a lightpath must be short, so it should cross few links.

The routing algorithms that are defined in order to achieve the desired mentioned reasons are:

1) Static routing: the routing table, the support, and the update are manually done.

$$GSNR_{AB} = \frac{P_{out,A} G_o L_1 G_1 \ldots L_N G_N}{P_{ASE,0} L_1 G_1 \ldots L_N G_N + (P_{ASE,1} + P_{NLI,1}) L_2 G_2 \ldots L_N G_N + \ldots + (P_{ASE,N} + P_{NLI,N})}$$

2) Dynamic routing: routing table, maintenance and updating runs by routing protocol.

3) Distance vector: according to the hop counts.

4) Link state: according to the state of the link.

3.2.1 Weighted graph

The key aspects in the design of optical transport networks are diverse, the most important are to be able yo support mesh topologies, having a dynamic allocation, an automated network control and light path setup.

A mesh networking means that each node in the network will behave like a hope that gives facilities for a lightpath propagation rather that being a blocking point in the road. For that reason the ROADM is introduced in a meshed optical network topology.

In such a meshed topology, the graph is considered weighted graph in which each edge element of the graph has specific positive value called "weight". Usually, the weight of a node refers to it's cost, that can be length of a route, the line capacity, the energy required to propagate along the routing space [10].
In order to specify the route of a lightpath having a source and a destination, the computed route is done by summing the weight of all the crossed nodes from source to the destination.

3.2.2 Shortest path algorithm:

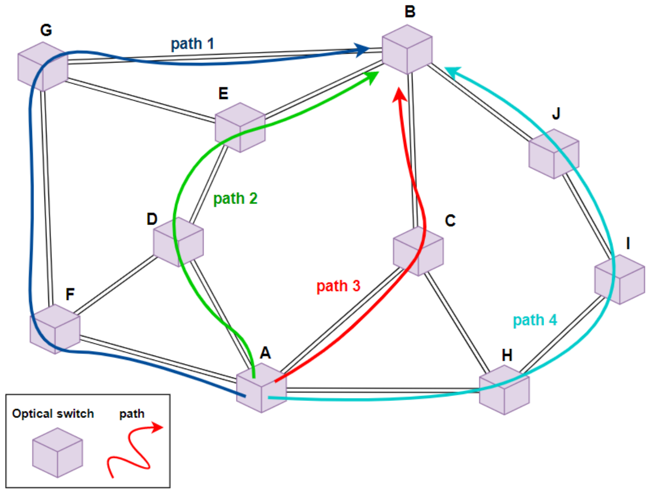

Figure 3.10: Example of shortest path algorithm

In order to find a path along the weighted graph mentioned before3.2.1, giving a source to destination end-to-end line, we run routing algorithms that take the decision on the chosen route.
The most considered routing algorithm in the networking domain is the shortest path algorithm. The graph topology done for a network is for the available paths from a node source to a node destination in which a node in an optical domain is a reconfigurable optical add-drop multiplexer (ROADM). Each edge represents a communication link as an optical line system (OLS).
This algorithm will choose among all the available paths which path has less number of hops to traverse and then this path will be allocated.

In the figure 3.10 above, node A sends a packet toward node B, the available paths are:

1) A-D-E-B
2) A-F-G-B
3) A-C-B
4) A-H-I-J-B

As we can see from a number of hops point of view, Path1 and Path2 traverse 2 hops, Path3 1 hop and Path4 crosses 3 hops.

path2 has the minimum and shortest-distance than all the other paths, which means path2 has the lower cost and will be obviously chosen as a winner of shortest path.

Moreover, the bandwidth demand driven by the internet and video traffic has grown very fast and as consequence the network operators pushed for the lowest network cost as a key point more than the performance [11].

3.2.3 Lightpaths over optical line systems:

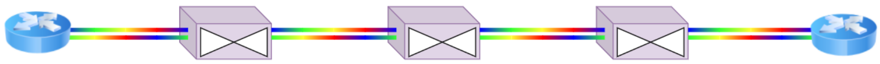

Figure 3.11: Optical line systems

The lightpath (LP) is a transparent route (or path) over the network at a given λ defined by switching matrices of the switching nodes and is a dedicated optical channel from source to destination.
LPs are dynamically allocated which allows the WDM to be implemented in a virtual way with the application programming interface (API) to independently manage the network subsystems.

An optical line system (OLS) is the combination of fiber and amplifier in which they are managed as a single network unit and they are considered bidirectional and symmetric.
At the input of an OLS there exists a set of WDM channels distributed on the transmission bandwidth according to the WDM grid Df.

In general, it is a point-to-point optical transparent bidirectional link 3.11 connecting two switching nodes considered as ROADM nodes, and made of fiber spans pairs. These fibers are transmission fiber lines in which they are periodically amplified by in-line amplifiers; two fibers one for each direction furthermore in-line amplifiers each is made of two amplifiers).

In an OLS, the Booster/preamp amplifiers are amplifiers at the output/input of switching nodes.
These amplifiers are key enabling technology for all optical networks in which the Erbium-Doped Fiber Amplifier (EDFA) is an optical amplifier used in the C-band. The gain bandwidth in C-band defines the transmission bandwidth in most of the networks. Typically, boosters and preamps are integrated with the switching node.

3.3 Digital twin in optical communication systems

The digital twin (DT) has been used largely in multiple fields. From a point of view of the technical challenges faced in optical communication, DT has the potential to provide a low-cost and lightweight solutions in order to address the wide challenges, and it has a fully broad application in the optical communication domain.
The introduction of DT to optical communication is used to apply multiple functions. These applications include the fault management, hardware configuration, and the transmission simulation.
The DT technology is applied in optical communication in order to ensure the safety and the stability of the operation of optical communication systems, optimizing the hardware and the network resource efficiency. It provides a dynamic monitoring and capability of numerical analysis of the transmission system.
The designed framework of DT for optical communication is composed of application layer, data layer,physical layer and model layer. The physical layer in optical communication system, contains various optical equipment, network elements, transmission modules and fiber links.
In general, inside an optical communication, right from the physical layer to the network layer, we expect a significant amounts of generated data from a wide range of sources. These data include a real-time update data according to the network status, the operating state of the equipment, and the performance of the transmission system.

However, using virtual models, the DT establishes a real-time connection between the digital and physical worlds through. It describes the physical state by applying a technological module, simulating the operation process, forecasting the tendency of changes, and optimizing the object performance.

In fact, the optical signal data contain the entire information from the transmitter up to the receiver. These information involve characteristic of the channel, the signal quality, and the performance of the transmission process. The optical transmission systems contains in the physical space, optical transmitters and optical channels, optical amplifiers, and receivers.
Instead, in the digital space, the simulation process of the transmission describes the entire process of optical signal transmission starting from the transmitter, going through fiber links. All is done after relay amplifications to the receiver [12].

3.3.1 GnPy model

The Telecom Infra Project (TIP) is a is a global community of engineering-focused companies, organizations and startups working on an initiative to develop and accelerate the deployment of open, disaggregated, and standards-based technology solutions [13].
Within TIP, the Open-Optical Packet-Transport Group (OOPT) is working on the adoption on a large scale of white-box models in optical networks for high quality connectivity [14].

Figure 3.12: Telecom Infra Project network

The deployment has been open up, in order to have the ability to predict the performances of an optical line system based on an accurate information of the optical transmission system. The OOPT environment exploits the Physical Simulation Environment (PSE) in order to provide an Open-Source developed software for planning, analysis and optimization. Its representation in Python is the GNPy (Gaussian Noise model in Python) [15].

Gnpy is used for QoT estimation and it successful because it is aware of the physical layer inter-networking. the GNPy exploits nodes input spectrum information accumulation on the decision signal, in order to calculate all the amount of the noise generated by the amplifier as the Amplifier Spontaneous Emission Noise (ASE) and the non linear interference generated by the fiber NLI. As already mentioned in this chapter, the total amount of NLI and ASE noises are the main elements to determine the generalized signal to noise ratio GSNR starting by a specific power per channel

$$GSNR = \frac{Pch}{(PASE + PNLI)} \tag{3.18}$$

The GSNR is the unique quality of transmission (QoT) figure for lightpaths, depends only on the spectral load and on the deployed equipment. Thus, the GSNR is used as a QoT meter in many heterogeneous environments for planning and controlling purposes, provided that it can be easily evaluated from common APIs running on network elements.
The input parameters for the GnPy model include a network topology as a json file described as a directed graph between network elements (Transceiver, Amplifier, Roadm and Fiber) in which each element is listed with its parameters [16].
In our work the json file related to fiber spans contains the fiber configuration, the simulation configuration and the spectral configuration, as in figure **??**.

For this computation, some vendors engineering make use of the transparency in which the same channel power (or spectrum power density) is forced to be the same in all spans regardless of their length/loss which is a sub-optimal strategy because ASE and NLI noise contributions should be balanced in each span with the launched power, but it brings simplicity in the design and the network configuration [16, 17, 18].

Chapter 4

Analysis

4.1 Optical Network operating system

Nowadays, the big data-centers as Facebook, google, Amazon have been found a strategy to achieve very high efficiency while having a reduction in the cost by introducing the disaggregation of the software from the hardware.
This disaggregation is based on the use of Transport API (TAPI)s decoupled from the operating system controller [19].

Our network architecture is based on a 'partially disaggregated' optical multi-domain, an OLS multi-layered WDM to control and management. Hence, the amplified lines connecting ROADMs may be independent WDM OLSs [3].
It is composed of multi-vendor OLSs while the transceivers use the WDM technology by introducing the ROADMs along a spectrum path in which we have continuous optical spectrum between the edge-nodes in the WDM layer.

The SDN controller system orchestrates the OLS transceivers and it relies on an optical SDN controller.
The interface between OPTICAL SDN and the OLS domain is based on the TAPIs, in which their contest topology is represented as nodes and links and can expose the internal topology. The OLS controller has an overall view of the topology, and exposes the real abstract topology for the scalability needed in an optical topology [20].It computes the end-to-end lightpath dedicated path computation element (L-PCE) between each source and destination point by exploiting the internal TAPI. Then, after having the list of connections for a path from the source to the destination, it implements the routing and spectral assignment algorithm for each propagated lightpath.

By requesting an optical signal, the OLS controller must provide the WDM connection between the involved transponders. The fist step is done by the optical network controller (ONC) that specifies the available routes and wavelengths in the system to be able to propagate that transparent LP. This propagation is done after running the routing and wavelength assignment (RWA) algorithm to select the links and allocates available cores.

After choosing a connection between a source city 's', toward a destination city 'd', in the considered transparent infrastructure, the L-PCE has a necessary need of the QoT-E in order to negotiate the topological graph that describes the network WDM layer and supported flexible grid in the C-band.
Each of the allocated WDM connections network domain, contains ROADMs based on wavelength selective switches with EDFA, 80 km of SSMF transmission line and an SDN controller (OLS controller) in which each transponder is equipped with a C-band modulated wavelength.

4.2 Optical Network analysis

4.2.1 Montecarlo Method

The Montecarlo simulations are used by multiple variable domains, including engineering, finance, supply chain, and science.
The basic idea is to assign multiple variety of values to an uncertain variable, in order to achieve multiple outcomes. The second step is to compute the mean averaged over the generated data to obtain a clear estimation of that variable.

In telecommunications and networking engineering, the Montecarlo analysis is used to evaluate and estimate network performance (in our case GSNR) in different variable scenarios (connection nodes in a network). Hence, the aim is always to optimize the network performance under an optimal point defined by each network scenario. It can be considered as a predictable outcome, hard to see in advance sometimes. Starting from the idea that when we increase the number of runs for a simulation, we initiate the development of the skeleton and the type of the distribution assigned to our results. The main two parameters while running a Montecarlo are the equation for the evaluation and the random variables for the input.

The main library of math calculation in python is called *Numpy* and is used with *Pandas* library to construct the model and create the generation of multiple results to study them in a relatively easy *'five-finger exercise'*.

4.2.2 Fundamental network architecture set-up under consideration

The disaggregated re-configurable optical add-drop multiplexers (ROADMS) and switches are at the helm of the responsibility to address the available channels wavelengths, according to the switching matrix, from a given input OLS to a given output OLS.
These ROADM switches, define the spectral load at the input of each line system in the abstraction of the network. The assigned software must accumulate metrics and then properly propagate signal information lightpaths in the routing space [3].

Our network topology is constructed by ROADMS on WDM C-BAND, connected by independent OLSs that include : the degrees of the ROADM multiplexers and demultiplexers, fibers of a single type "SSMF" and the same transmission equipment with a dispersion D= 0.5 10^{-5}. It contains input connectors right after the fibers, and amplifiers "EDFA" (booster, inline and preamp) as shown in figure 4.1. These OLSs transport lightpaths WDM C-BAND optical signals are considered as fully loaded system from ROADM to ROADM independently. Hence, every channel propagated along the OLS is uncorrelated with the other channels propagated contemporary.
Additionally, we focus our study on the channel under test ("cut" channel). Hence, when we don't have Raman effect, this channel becomes the representative case that represent the most of the non-linearity as a worst GSNR.

The following figure 4.1 present the architecture of an OLS in our network topology:

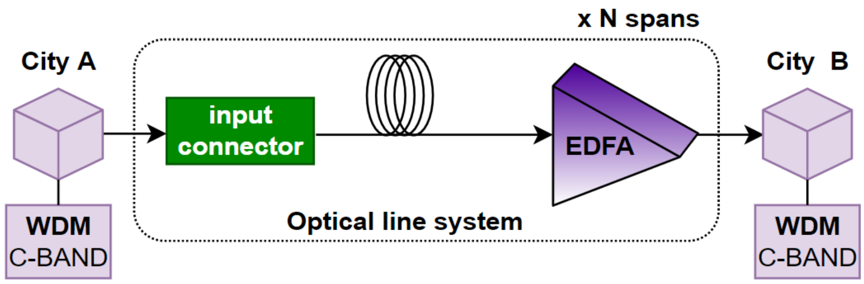

Figure 4.1: Conceptual schema of a ROADM-to-ROADM WDM optical line system

In a network topology, the problem of having a uncertainty on the input parameters that directly affect the QoT remains an essential key that we study in our system.

Instead of considering a constant input connector loss, constant ripples , and a unique length of connections, we work on a real topology that collect in the routing space multiple node to node connection each of different length, with random input of connector loss and ripples.
The length of each connection between ROADMs is collected according to the real distance in a real topology of Deutsche Telekom (DT) of German shown by the following figure 4.2.

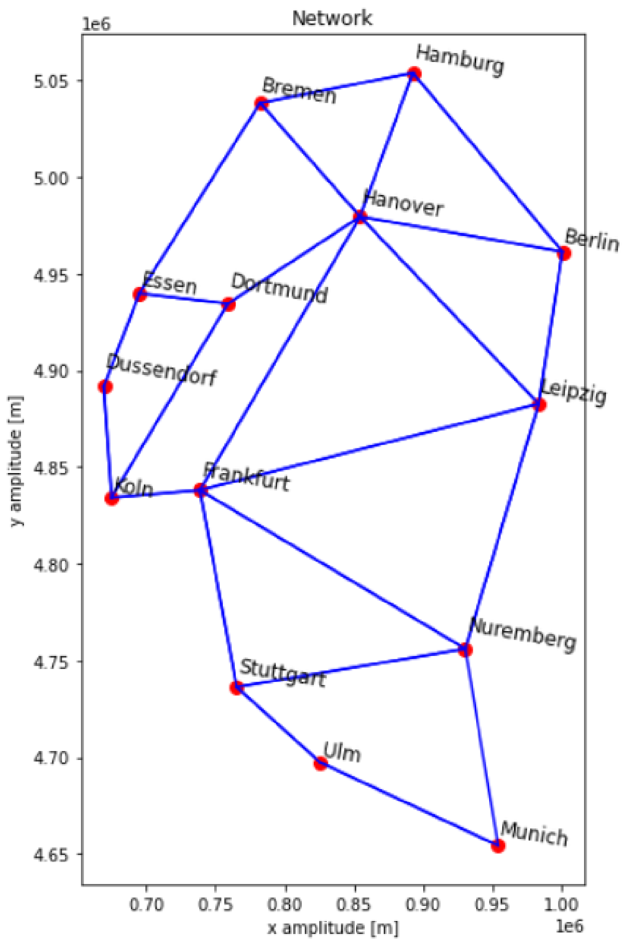

Figure 4.2: German topology DT from python

Moreover, as shown in figure 4.2 we collect the lengths for OLS between each couple of connected cities, then we introduce a fiber span for each sub-path of tranche of length varying between 50 km and 100 km, knowing that the NLI effects are always created at the first 30 km of the OLS.

Then, in our python code, for each lightpath propagated, after computing the numberN_{fiber} spans that are needed for the corresponding length of connection, we assign $N_{amplifiers}$ for that OLS in order to compensate for all the power loss caused by the fibers.

In general, the operator in real systems have uncertainties, and it doesn't have the real description of the connector input loss, ripples and fiber span losses at the input of each OLS.

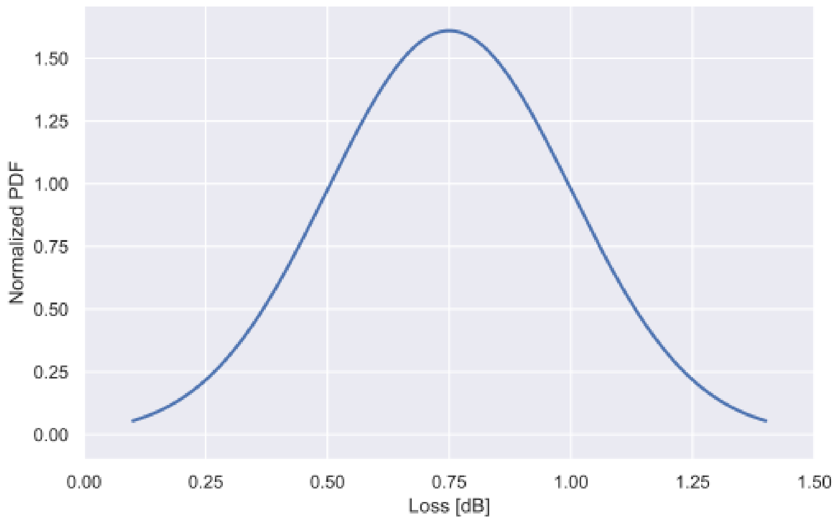

Figure 4.3: input connector loss pdf

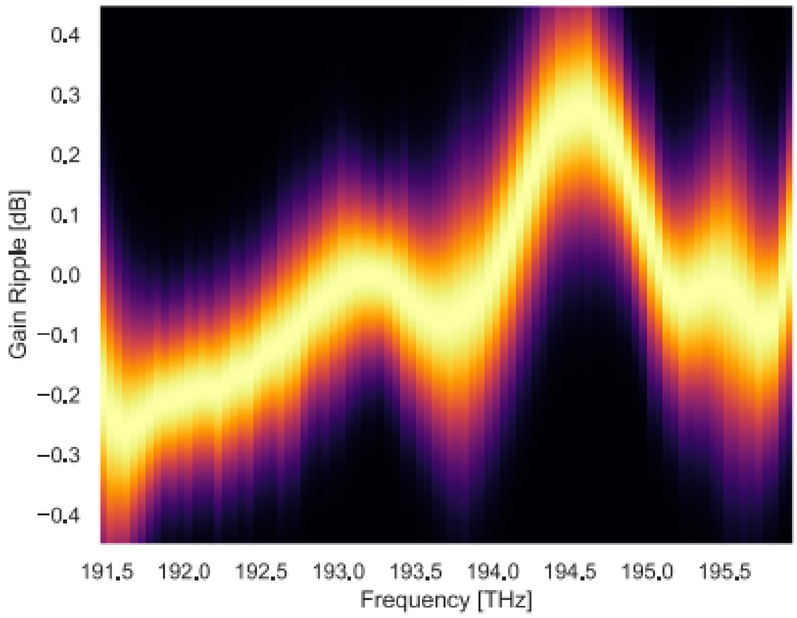

Figure 4.4: Heat-map of the normalized gain ripple profile PDF vs. frequency

We generate the input connector loss following a Gaussian distribution as shown in figure 4.3 in which we start by picking randomly choice of loss to the first span, then we pick another choice to assign it for the second and so on up to the end of the OLS. Hence, all these parameters are statistically independents.
Then, we assign for each of these losses a random Gaussian distribution of the ripples added by the amplifiers in which each vertical line in the ripples 4.4 is a Gaussian distribution around the mean.
The propagation of lightpaths in our system is done in transparency, which means we have the same fiber type along all the OLS end-to-end. Therefore, we assign the same power at the input of all the fibers in line.

In real lines, after each amplifier there is a connector with the next fiber. The output connector doesn't affect the non-linearity, so we neglect it's effect, because NLI depends on the power per channel P_{ch}. The NLI is created at the end of the propagation of the fiber span. In order to obtain the real NLI at the end of the fiber span, we have to multiply the NLI by the loss fiber due to fiber attenuation. Then, ASE noise is also generated by each amplifier trying to compensate to the NLI noise generated by the fiber span at the input in order to have the exact input power at the output destination port.
The amplifier read the total power at the input and at the output. So after all the

added gain, if the end power is not the desired one, it applies a shift.

Hence, the implementation of the propagation is done in order to obtain after the EDFA the same input power before the input connector even if it was a random input.
The amplifiers are provided to compensate for signal attenuation, When the information signals are transmitted over long distances, or between links of optical fiber cable.
Therefore, according to the losses, we set the gain to every amplifier and then all the amplifiers are equipped by input power and output recovered compensated power.

The power per channel must be sufficient to provide an adequate GSNR in the presence of the ASE noise from the amplifiers, necessitating a high amplifier total output power for systems with high fully-loaded capacity.
The launch power per channel P_{ch} into each span should ideally be re-optimized according to the changing WDM comb characteristics. This function is typically attributed to a "physical-layer aware" Control Plane (CP). Basing on the future state, the overall network should automatically optimize all values of P_{ch}[21] after a reconfiguration.
In order to set the optimal power for each input fiber, we start from the last amplifier. The signals arrive to a node with a fixed power, in our case -1dB. Then, the last amplifier should have a gain that bring the signal to the original propagated power = -1dB.
Knowing the loss of the last fiber, we set the optimal power at the end port for the last amplifier, then we proceed backward in order to recover the input power.

4.2.3 Experimental simulated method

A transparent optical network is considered as a meshed weighted topology where ROADMs nodes and switches are connected by OLSs. Thanks to the capability of switches to address any wavelength from any source to destination, wavelengths propagates transparently, over the corresponding lightpath. These wavelengths are impaired by the crossed OLSs and the in line switching nodes.

In order to have a real efficient description of a WDM optical transport from a data structure, we study the topology of weighted graph, in which the graph edges are the OLSs and the nodes are switches and ROADMs.

Starting from the network of German DT 4.2 having 14 cities connected according to a meshed weighted topology graph considering as a weight the GSNR corresponding to each city-to-city line in which we assign for them a dataset.

The weight of nodes and edges are main description for the QoT in which the overall merit of a lightpath is summarized by the "GSNR".

Using the algorithm of finding the path from a source to a destination city based on the 'shortest-path' and the minimum number of crossed nodes (hopes), we propagate a high number of independent lightpaths by creating connections based on the Montecarlo approach between a source Roadm and a destination ROADM. Accordingly, it collects the number of fiber spans needed, the starting and the end city name.
All the other possible paths from the source city to the destination are not considered.
In order to perform our study analysis, we create another dataset based on Montecarlo runs for all the propagated connections, containing the information needed for the QoT-E (optimal power, GSNR mean and standard deviation).

The QoT-E must be used in order to compute the GSNR of a transparent LPs to assess network performance before, after and during deployment. This estimation is done by enabling an implementation of the lightpath computation engine (L-PCE).

In this work we perform an experimental validation of the consistency between GSNR calculated by prediction along an OLS from ROADM input to a ROADM output methodology, and the GSNR calculated by an analytical method as an optimal conservative GSNR.

Our assumption is done by considering the operation of LPs as additive white Gaussian noise (AWGN) channels. After collecting the set of propagated connections, we continue our calculation by running the Montecarlo of 10,000 number of runs for each single connection of variable input connector loss and EDFA gain and ripple. Then, we measure the NLI and the ASE for a set of input powers, each with distinct gain ripple and loss respectively that follows a Gaussian distribution. Our approach takes into account as nonlinear effects NLI the SPM only without including the XPM, FWM and the Raman effects.

Thanks to the relation between P_{ch}, ASE and NLI, we compute the predicted GSNR. This GSNR is obtained by the study done along each OLS for the corresponding lightpath connection in a point-to-point optical amplified line, containing N erbium-doped fiber amplifiers and N fibers. The GSNR is considered as a first analysis to search for a common statistical characterization.

The number N of fibers is chosen by our python code according to the total length of the connection allocating a fiber followed by an amplifier each 80 km, for example for a connection from Ulm to Munich:

$$N_{fibers} = \frac{distance_{(Ulm,Munich)}km}{80} \tag{4.1}$$

We obtain at the end 10,000 values for each of GSNR, OSNR and SNR_{NL} according to the values generated by Montecarlo for each propagated lightpath.

4.2.4 Analytical method

In addition to the approach done experimentally, we implement an analytical model on top of all the created connections for the same network topology in order to predict the GSNR optimal, and for comparing the estimated end-of-line GSNR degradation to real QoT measurements done by the system.

What we do is, for each city connection we take the entire path with all the hops traversed by the transparent LP, then we split for each consecutive couple of cities source "s" and destination "d". Each couple is treated aside taking data from our dataset of node-to-node connections. This dataset contains information for mean and standard deviation of GSNR on the corresponding optimal power, following the same procedure done on top of the 500 full line connections computations.

The prediction on the mean of the total GSNR end-to-end line connection is collected as:

$$GSNR_{s,d} = \frac{1}{\sum_{i,j}^{OLS} \frac{1}{GSNR_{i,j}}} \tag{4.2}$$

While the standard deviation is predicted as the sum of normally distributed independent variables as these sub-couple-paths that are statistically uncorrelated. Hence, these independent random variables are normally distributed, then their sum is also normally distributed.
The total standard deviation along the OLS is computed as:

$$std_{s,d} = \sqrt{\sum_{i,j} {std_{i,j}}^2} \tag{4.3}$$

where i,j are labels that represents the ROADM OLS source and destination crossed by the LP, while $GSNR_{i,j}$ is refers in turn to the respective GSNRs for the specified wavelength of the channel under test (CUT), λ_{CUT}. The spectral load and the OLS model are provided to the QoT-E by the ONC. The original QoT provided is the $GSNR_{i,j}$ that is expressed as:

$$GSNR_{i,j} = \frac{1}{\frac{1}{OSNR_{i,j}} + \frac{1}{SNR_{i,j}}} \tag{4.4}$$

$$OSNR_{i,j} = \frac{P_{CUT,ij}}{P_{ASE,ij}} \tag{4.5}$$

$$SNR_{NL,ij} = \frac{P_{CUT,ij}}{P_{NLI,ij}} \tag{4.6}$$

These QoTs parameters are the optical SNR that are collected from a ROADM source city "s" to a ROADM destination city "d". They are accumulated thanks to the knowledge of the ASE noise, $P_{ASE,ij}$, and the NLI interference, $P_{NLI,ij}$, respectively.
However, $P_{CUT,j}$ is the power of the channel under test that we input at the beginning of the destination ROADM of for a specific OLS_{ij}, and surely the power considered is the power optimal.

4.3 Gaussian distribution normality Tests:

In statistics and data science, hypothesis testing and statistical inference relies heavily on the normal distribution. Hence, what we look at in order to do our study, firstly, to prove that our generated distribution are normal then using the first and the second mode of a distribution: Mean μ and Standard deviation σ. Then, in order to obtain a precise estimation of the total GSNR within the OLS under consideration, it is enough to compute for GSNR μ_{GSNR} and σ_{GSNR}.

We base our assumptions on the The Central Limit theorem (CLT) which says that if we sum an infinite number of a distribution, no matter what is the distribution of each variable, we should obtain a Gaussian distribution. This normality is obtained in case the variables are statistically uncorrelated [22].

In this section, our main interest is falling on proving that our data samples of GSNR of size 10,000, follows a normal distribution by using multiple graphical statistical tests.
The graphical test are based on plotting the generated dataset to check if the behavior is consistent with a bell-curve, the extended for a more robust graphical check with the Q-Q method.

Then, the analytical tests assume that the sample is drawn from a Gaussian distribution following the so-called null hypothesis(H0). We prove this assumption by putting a threshold level that we compare to an output called p-value returned by the test to accept or reject the assumption H0.

1) H_0 = The sample comes from a normal distribution.

2) H_1 = The sample is not coming from normal distribution.

The normality of our generated data is a basic assumption that we seek to obtain. It is defined by the density of probability [23]:

$$f = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.7}$$

Where μ is the mean value, σ standard deviation of the corresponding dataset x. We test the normality by two basic tests of normality: Test of "Shapiro-Wilk" and "D'Agostino's K^2 Test". We use multiple prove for the normality because the failure of one normality test means that your data is not normal [24].

4.3.1 Shapiro-Wilk Test

In practice, the Shapiro-Wilk test is believed to be a reliable test of normality. It evaluates a data sample and quantifies how likely it is that the data was drawn from a Gaussian distribution. Shapiro-Wilk test returns a critical value and the p-value that should be compared to a threshold called α which is typically equal to α= 0.05. Anyway, α is the probability of rejecting the null hypothesis H0 when it is true.

1) p $\leq \alpha$: reject H_0=> not normal.

2) p $> \alpha$: fail to reject H_0 => normal.

What we seek for is to obtain a p-value larger than α enough to accept the assumption that our data samples are normally distributed [25].

In python, the shapiro() SciPy function calculates the Shapiro-Wilk on a given dataset as shown by the code below:

```python
# Shapiro-Wilk Test

from scipy.stats import shapiro

# normality test
stat, p = shapiro(data)
# interpret
alpha = 0.05
#initialize the normal variable by False
Normal=False
#update when p> alpha
if p > alpha:
    Normal=True
```

Having the null hypothesis H0 is the hypothesis that the dataset is normally distributed. The Critical values in a statistical test are a pre-defined range significance boundaries at which H0 fail to be rejected if the calculated statistic is less than the critical value.

We can interpret the results by failing to reject the hypothesis H0 and that the data is normal in case the calculated test statistic is less than the critical value at a chosen significance level.

4.3.2 D'Agostino's K^2 Test

Agostino K^2 test is our third normality check, and one of the most important. The summary calculated by this test of the statistics from the data are called Krtosis and skewness[26]. However, our casual interest is to determine if the data distribution departs from the normal distribution.

1) Skew: it is a measure of asymmetry in the distribution. It quantifies how much a distribution is pushed left or right.
2) Kurtosis: It is a statistical test for normality, it quantifies how much of the distribution falls in the tail.

This test compare also the critical value α to a p-value calcuated by the test, in which we expect for the normality that p $> \alpha$ [27].

The D'Agostino's test can be applied via the normal test() SciPy function in python, and it returns the test statistic with the p-value as shown in the following code:

```
# D'Agostino and Pearson's Test
from scipy.stats import normaltest
# normality test
stat, p = normaltest(data)
# interpret
alpha = 0.05
#initialize the normal variable by False
Normal=False
#update when p> alpha
if p > alpha:
    Normal=True
```

4.4 Statistical analysis:

In general, an analytical accumulation analysis is a set of statistical methods that we use in order to estimate the relationships between a dependent variable and one or more independent variables. These analysis are used to demonstrate the strength of the relationship between variables and to model the future relationship between them [28].
In our activities, we use statistical analysis to prove the relationship between our QoT simulated GSNR on a full path with an analytical separated independent variable coming from the accumulation of sub-paths of the same path 4.2.

4.4.1 Comparing Distributions: Z-test

Z-test is one of the most important tests in statics that is used to compare the distribution of two samples and conclude if both of them follow the same normal distribution. The purpose of this test is to see if two distributions are significantly different from one another. In general, Z-test performs as a hypothesis test in which the z-statistic follows a normal distribution.

In addition, in this test, comparing two Samples dataset means finding the difference of the two sample means in units of sample mean errors. This difference in terms of significance is computed as Z-significant [29]:

$$Z = \frac{(\overline{X_1} - \overline{X_{x2}})}{\sqrt{\sigma_{x1}^2 + \sigma_2^2}} \tag{4.8}$$

$$\sigma_{x1} = \frac{\sigma_1}{\sqrt{N1}} \tag{4.9}$$

$$\sigma_{x2} = \frac{\sigma_2}{\sqrt{N2}} \tag{4.10}$$

$\overline{X_1}$ and $\overline{X_2}$ are the averaged mean of data sample 1 and data sample 2 respectively. Meanwhile, σ_{x1} and σ_{x2} are the standard deviation of each of sample 1 and 2 divided by the size of each sample N1 and N2 respectively.
We suppose to have these assumptions:

1) H_0: Both dataset follow the same normal distributions
2) H_1: The two distributions are different.

We apply the two-tailed test that by convention is used to determine significance at the 5% level, which means each side of the distribution is cut at 2.5%.

Our rejection region is : Z <= - Z2.5 and Z>=Z2.5 (assuming 5% significance level, split 2.5 each on either side).
This calculated Z with a 95% probability will fall between the two limits defined by: - Z2.5 = -1.96 and Z2.5 = 1.96 as a confidence interval[30].

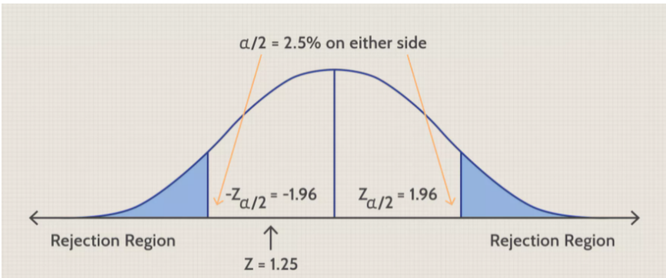

Figure 4.5: Z test using two-tailed test rejection region

4.4.2 Extreme conservative margin with Euclidien distance

Three-sigma limits

The Three-sigma limits also called the Empirical rule is an engineering quick estimating tool of the spread of data in a normal distribution. This test is done by considering a statistical calculation in which the data under test are within three standard deviations σ from a mean μ. Graphically, On a bell curve, it gather all data that lie above the average and beyond the three-sigma line represent less than 1% of all data points.

The steps to follow in order to do the Three-sigma limits test [31]:

1) First, calculate the mean μ of the observed data.

2) Second, calculate the standard deviation σ of the set.

3) Finally, calculate three-sigma ($3\sigma_{GSNR}$), which is three standard deviations more or less the mean μ.

This test predicts that 99% of observations falls within the first three standard deviations of the mean ($\mu_{GSNR} \pm 3\sigma_{GSNR}$), thus, 1% lies outside of three standard deviations. In general, the Empirical rule is used in statistics for forecasting especially when obtaining the exact data is difficult.

Generally, in an optical network, vendors provide nominal values for the operational parameters of each network element (NE) that personalize the physical layer. NEs experience a modification in the working point that insert an uncertainty from the nominal value, which implies an uncertainty in the GSNR computation and requires the deployment of a system margin.
Also, due to hardware aging and to the change is the OLS spectral load, the equipment installed is affected by a fluctuation between the operational and the nominal point. This variation makes the simulated GSNR be different with respect to the nominal value computed by the QoT-E.

In our system, we have uncertainty on the ripples and the input connector loss, and that leads us to set a margin on top of our computed GSNR nominal on a LP in order to avoid out-of-service (OOS) events.
For each generated LP we establish the lowest possible margin required for the OoT-E. Then we apply the Empirical test to set an optimal margin for our predicted GSNRs by applying the $3{*}\sigma_{GSNR}$ below the nominal GSNR mean that we estimated, in order to avoid the system to run out of service while propagating a GSNR smaller than the needed one.
This margin is applied to set the maximum out of service probability in our network topology.

Euclidean distance

Furthermore, we check how much our predictions are far from our simulated data after applying the margin of conservation on the predicted values. Besides, we apply the Euclidean distance in order to calculate the Root Mean Square Error (RMSE) .

The Euclidean distance and the RMSE are represented in the following formula:

$$d\left(GSNR_{SIMULATED}, GSNR_{optimal}\right) = \sqrt{\sum_{i=1}^{n}\left(GSNR_{SIMULATED,i} - GSNR_{optimal,i}\right)^2} \tag{4.11}$$

$$RMSE = \frac{\sqrt{d\left(GSNR_{predicted}, GSNR_{actual}\right)}}{N} \tag{4.12}$$

in which N is the size of our data sample under test. This method is characterized by the ability to give the facility to eliminate the methods that have the most significant errors [32].

4.5 Transfer learning application for QoT-E

The transfer learning techniques are mainly deployed in order to have a method to decrease the uncertainty that we face on the GSNR computation and to arrive for a minimum applied margin of conservation.

In optical networks, the ML is a paradigm that is expertly applied to monitor the performance of the network. For this reason, an algorithm is introduced to perfectly exploit the available database to design margins and reduce the uncertainties in the network [3].

The neural network model applied is deep neural network (DNN) and is trained to predict the QoT-E in term of GSNR before the deployment of an optical LP from the source to the destination.

In our activity, we consider two networks characterized by different topologies, German topology and the USNetwork topology. These two networks have the same transmission equipment(NF, chromatic dispersion and frequency), fiber type (SSMF) and amplifier type (EDFA).

Then, we initiate by training a ML agent running on top of the dataset generated for the known German network topology. In general, as the uncertainty in a LP QoT is not realistic, the ML scope is to be able to correct the GSNR computation in order to reduce this uncertainty due to EDFA ripples and the input connector losses.

The schematic of the ML model trained on the dataset of German network topology of 500 connections is implemented within the QoT-E of US network to be able to correct the prediction on the GSNR. This correction must be done before the deployment of the LP.

The main elements of ML model are the features with the label that we feed the network to do a transfer learning. As features, we consider the input connector loss, ripples and the number of fiber spans per connection, while we consider the GSNR as a label to be predicted.
In fact, our work is done basing on the open-source project Tenserflow library.

A dataset is created to train the ML based on the German topology, whereas a test dataset is generated from the US topology to test the efficiency of our ML model and to study the prediction of the GSNR as a QoT. With these fixed parameters, we guarantee that the input connector loss and the ripples are the same on the shared independent propagated sub-lines.
We train the network of exact powers regarding GSNR and then we test by transfer learning on the test dataset in US network.

Chapter 5

Simulation results

In this section we present the created dataset with the results obtained by applying the activities on top of them.
Firstly, we generated all the data related to the 500 randomly chosen connections for each city by running the first Montecarlo.

The essential elements remain to collect for the network are the GSNR, OSNR and SNR that mainly come from the calculation of the ASE and NLI noises along the OLS. These parameters are collected on the optimal power for the central channel "channel-cut" that is a WDM C-Band channel, in which all the channels are considered fully loaded in the network topology with no Raman effects and no dependency between the channels.
For that reason, We pass these information to another Montecarlo on a fixed OLS that is running on top of the uncertainties of the input connector loss and the ripples, and generating for each single connection 10,000 runs for each line connection by propagating a lightpath in a fast propagate method implemented in python. Then, this program do the collection for each run the ASE and the NLI per power propagated, picked from a power sweep.

5.1 Synthetic dataset generation

5.1.1 Dataset1: d1

This dataset is for the experimental analysis, it contains 500 different random generated connections from a given source to a destination city. This connection should always be chosen according to the shortest path. Then, the length of every chosen path is collected to find the number of fiber spans along the OLS.

We expect our generated dataset are normally distributed for the mean and standard destination following the Central Limit Theorem (CLT) concept of the many repetition of a Montecarlo model. As mentioned before, we collect the averaged mean and standard deviation of the GSNR over the total runs of the second Montecarlo, along the OLS, allocated between each couple of cities as an essential metric of study measurement. In fact, we collect the averaged mean of the QoT parameter. **Example from dataset1**

connection	N_{fibers}	$GSNR_{mean}$ [dB]	$GSNR_{std}$	OLS_{length} [km]
Bremen->Hannover->Frankfurt ->Stuttgart->Ulm	10	18.999	0.151	729.7
Leipzig->Numberg->Stuttgart	6	21.644	0.1	463.4
Berlin->Hannover-> Bremen	6	22.6348	0.0868	416.2

Table 5.1: Examples of connections in dataset1

5.1.2 Dataset0: d0

On the other hand, this dataset is for the analytical study. This dataset0 is a more smaller dataset that contain the connection between each couple cities connected by an independent OLS as shown in on our weighted graph. This means that each connection is made by only 2 node-to-node connected cities.
For each line, we run again the 10,000 Montecarlo algorithm in order to collect the same parameters for the QoT study. As done in dataset1, we collect for dataset0 GSNR, OSNR and SNR after inserting a specific input power and we work on the optimum one and the central cut-channel over a WDM C-Band.
Example from Dataset0

connection	N_{fibers}	$GSNR_{mean}$ [dB]	$GSNR_{std}$	OLS_{length}[km]
Essen->Bremen	4	25.67	0.05	278.5
Leipzig->Frankfurt	5	22.63	0.086	352.6
Berlin->Hannover	4	25.69	0.049	294.9

Table 5.2: Examples of connections in dataset0

5.2 Distributions of GSNR, OSNR and SNR for d0 and d1:

5.2.1 Normality tests:

In general, a large fraction of the field of statistics is concerned with data that follow a Gaussian distribution. In this section, we want to prove the CLT concept that explain the fact of repeating an event for a high number of runs gives a Gaussian distribution specifically for the GSNR, OSNR and SNR parameters in which each of them is a dataset of 10,000 values.

We start by showing graphical results in order to be able to see that a distribution is Gaussian by the bell-shaped curve histogram as a first step. Then, we plot the quantile–quantile (Q-Q) plot in order to have more accurate graphs that show the normality to be true beyond doubt.

In addition to the graphical show, we prove by an analytical approach to make the magic of obtaining a Gaussian shape be confirmed in reality by statistical computations.

5.2.2 Graphical normality check using Histogram

The graphical method is used for plotting the histogram of our generated data and qualitatively evaluate whether the data looks Gaussian as a bell-shaped curve [24]. We focus our study on estimating the GSNR over each connection and evaluating the alternation in the quality of the signal that happens when modifying the spectral load connection.
Given the different signal power dependencies of the ASE and NLI contributions, in an optimal working point scenario the former is the most significant contributor to the GSNR degradation, as it is twice the NLI. We present how every per-wavelength GSNR distribution can be completely approximated as Gaussian distributed.

From DT German topology, we choose randomly a connection for each of dataset1 and dataset0 and we plot the histogram of the distribution for GSNR, OSNR and SNR_{NL}. On top of each of these histogram distribution, we plot a real normal distribution by computing the mean and the standard deviation for the corresponding connection dataset.

connection between Berlin-> Numburg:

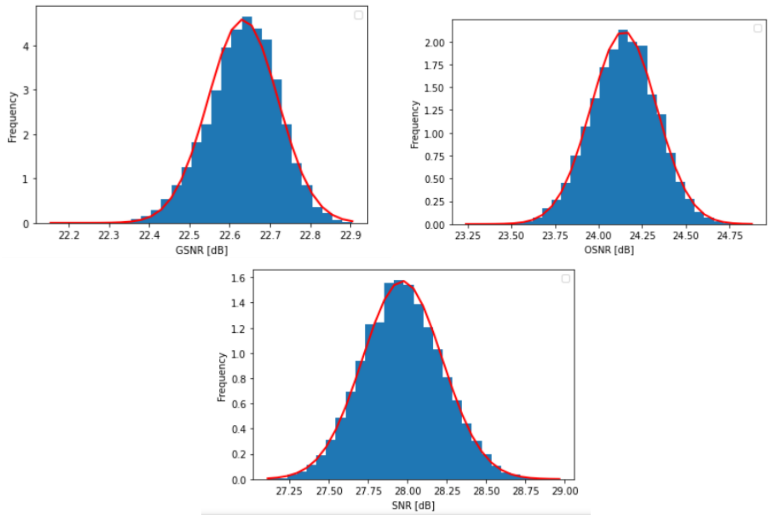

Figure 5.1: GSNR, OSNR and SNR_{NL} distribution from dataset1, Berlin-> Numburg

Connection between Frankfurt->Ulm:

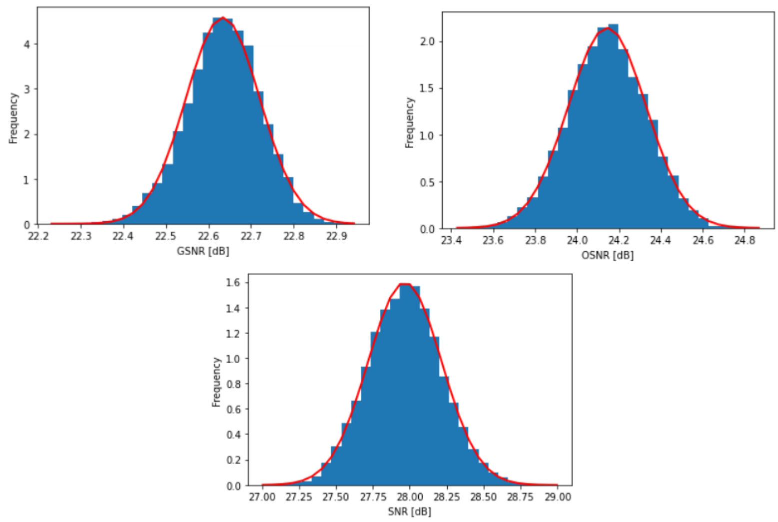

Figure 5.2: GSNR, OSNR and SNR_{NL} distribution from dataset1 Frankfurt ->Ulm

Connection between Essen->Bremen

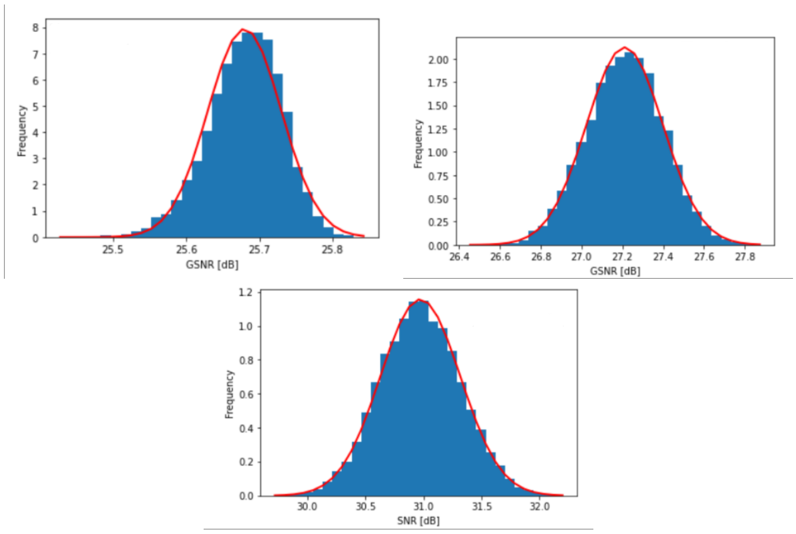

Figure 5.3: GSNR, OSNR and SNR_{NL} distribution from dataset D0 Essen ->Bremen

Connection between Ulm->Munich :

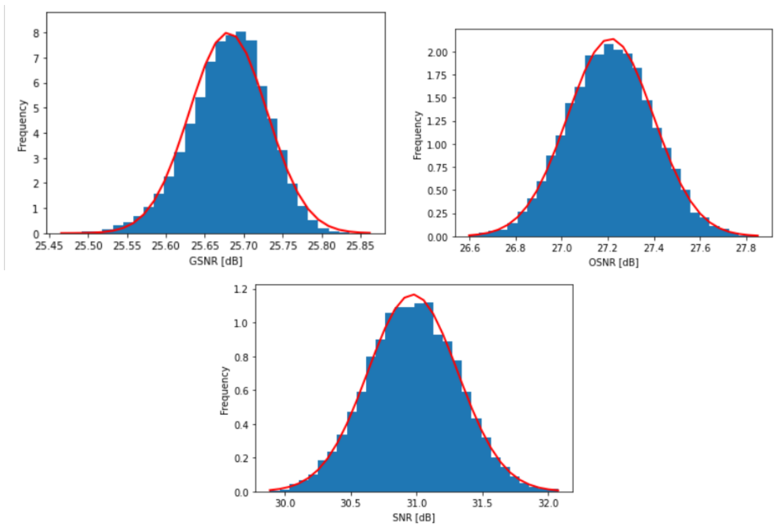

Figure 5.4: GSNR, OSNR and SNR_{NL} distribution from dataset D0 Ulm ->Munich

the figures above 5.4, 5.3, 5.2 and 5.1 clearly show that the distribution of GSNR, OSNR and SNR_{NL} for dataset1 and dataset0 is always shaped as a bull-curve, and they are distributed around the average nominal value according to a probability density function (PDF) that is well approximated as Gaussian. This result is in complete agreement with the Gaussian distribution and is centered around the normalized correspondent mean.

5.2.3 Graphical normality check using Q-Q plot

In some cases, the histogram plot may show a bell-shape curve, but in fact the distribution just looks normal instead they may come from a student distribution. In order to ensure our assumption, we trace the Q-Q plot. We assume already that our data are compared to a Gaussian distribution passing the line argument as 's' to the qqplot() statsmodels function python (qqplot(GSNRs, line='s')).

In general, Q-Q plots the data samples, sorts them in an ascending order, then plots them versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of our samples data[33].
Hence, the perfect match is presented by a line of dots on a 45° angle from the bottom left of the plot to the top right. Often, a line is drawn on the plot to help make this expectation clear. However, any dispersion by the dots from the line shows a deviation from the expected distribution.
We tried the Q-Q plot over all the generated GSNRs for every single connection and all of them showed a perfectly line of dots following the 45° angle line for the normal distribution. For simplicity, we present a couple of connections from dataset1 and dataset0.

Essen->Bremen

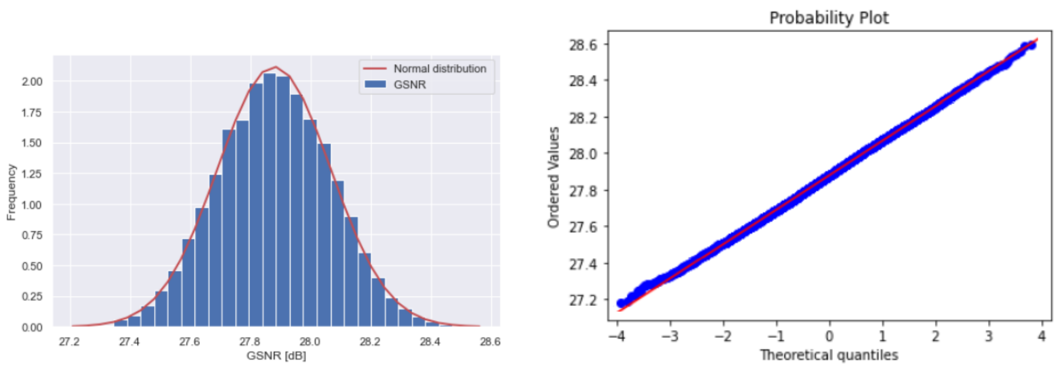

Figure 5.5: GSNR distribution from dataset0 Essen->Bremen

Stuttgart->Frankfurt

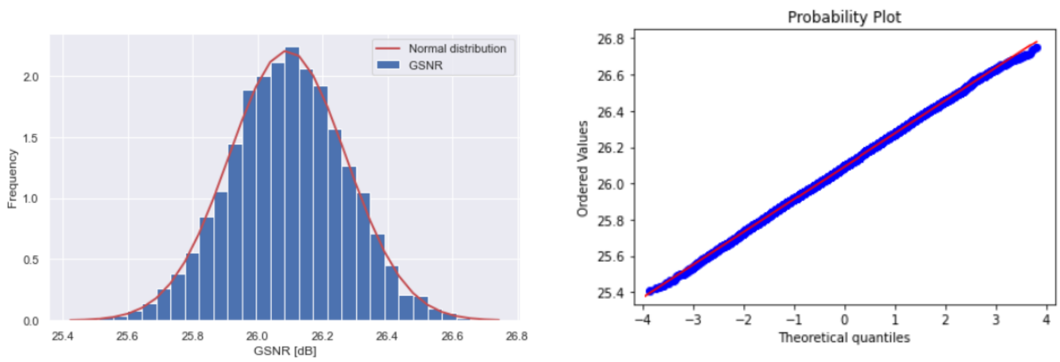

Figure 5.6: GSNR distribution from dataset0 Stuttgart->Frankfurt

Hannover ->Berlin

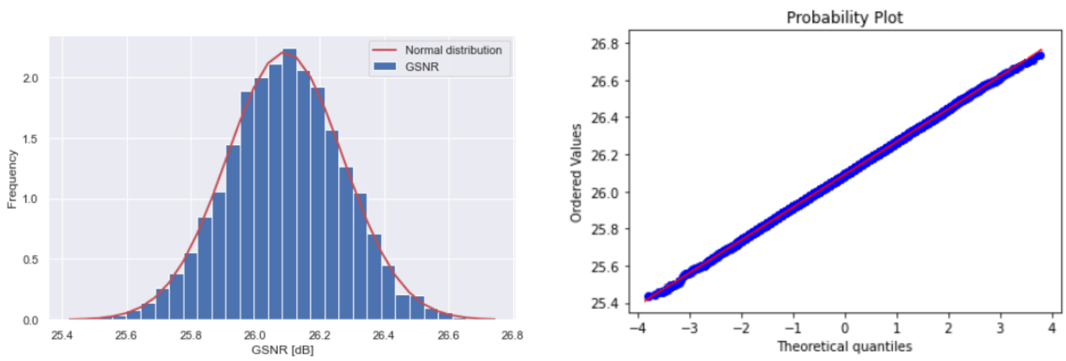

Figure 5.7: GSNR distribution from dataset0 Hannover ->Berlin

Essen ->Dortmund

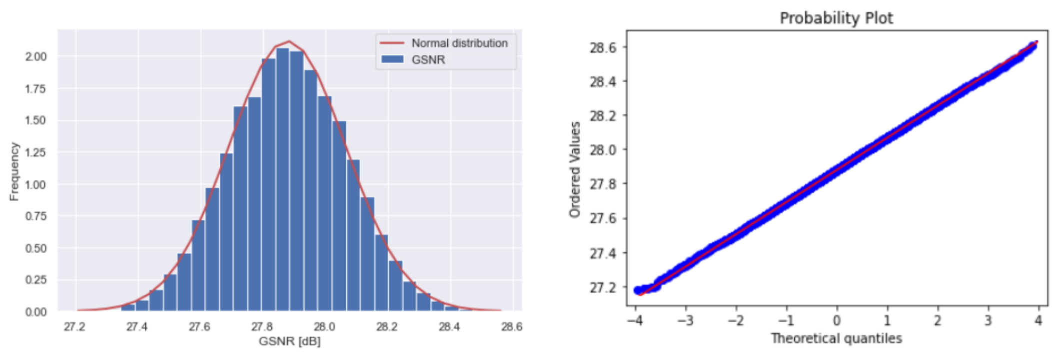

Figure 5.8: GSNR distribution from dataset0 Essen ->Dortmund

Essen ->Hannover

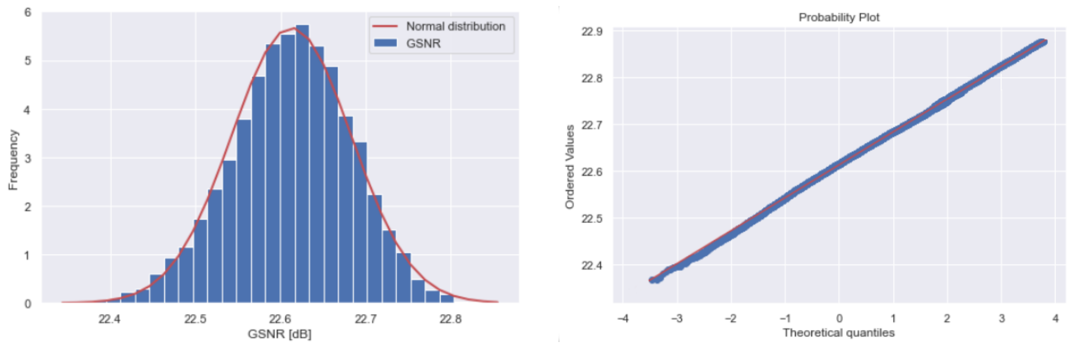

Figure 5.9: GSNR distribution from dataset1 Essen ->Hannover

Koln ->Berlin

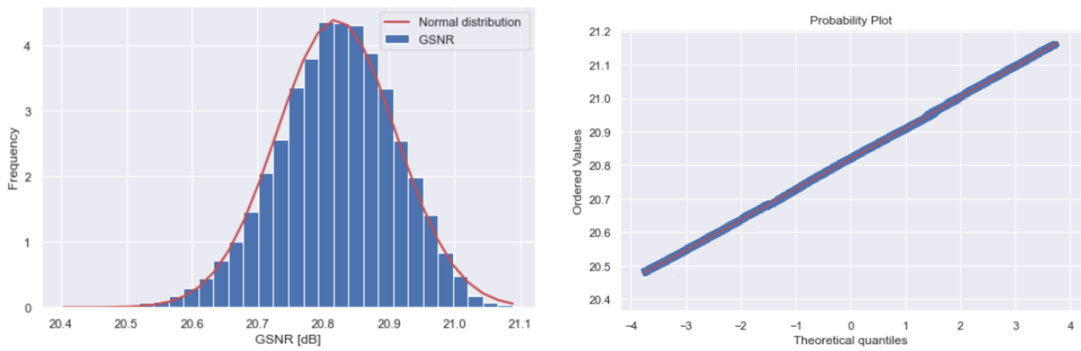

Figure 5.10: GSNR distribution from dataset1 Koln ->Berlin

As clearly shown by the figures plots above 5.10, 5.8, 5.7, 5.6, 5.5 and 5.9 examples between random couple of cities, all the quantile points lie along the red lines of 45° from the bottom of the graph in each connection. The quantiles represent our data sorted in ascending order, and they tent to dispose in the shape of a line almost entirely straight.

5.2.4 Statistical Tests:

In this section we present the statistical methods that we used on our dataset to quantify how likely the data was drawn from a Gaussian distribution.
Each of these tests looks at the normality from a different perspective, and like that we ensure the Gaussian distribution for our generated parameters.

Hence, these tests are applied on top of the whole dataset of 500 connections for each OLS connection, and clearly we prove statistically not only graphically that the distribution of GSNR follows a normal distribution .
In term of simplicity, we choose randomly paths from the whole dataset and we show the successive results related to each test of normality.

Shapiro-Wilk Test

In this test we generated for each GSNR the p-value in which we compare with the typical α value =0.05 instead of generating a list of critical values. The normality is present for the dataset GSNR in case p> α.

paths from Dataset1

paths	α	P-value	GSNR Normality check
Dusseldorf-> Leipzing	0.05	0.5	Normal
Hamburg-> Munich	0.05	0.54	Normal
Frankfurt->Ulm	0.05	0.3	Normal

Table 5.3: Normality check using Shapiro-Wilk Test for D1

paths from Dataset0

paths	α	P-value	GSNR Normality check
Essen->Bremen	0.050	0.550	Normal
Ulm->Munich	0.050	0.885	Normal
berlin->Hannover	0.050	0.415	Normal

Table 5.4: Normality check using Shapiro-Wilk Test for D0

In general, we present the results of this test by it's mean and standard deviation. The test presented successful output of normality for all the connections in dataset1 and dataset0. To present the results we report the mean and the standard deviation of the p-values over all the connections propagated in our dataset0 and 1.

Dataset	$\mu_{P-value}$	$\sigma_{P-value}$
D0	0.447	0.140
D1	0.427	0.11

Table 5.5: Normality check using Shapiro-Wilk

As shown by table 5.5 the mean of p-value for both dataset is greater than the threshold α 0.05 with a standard deviation around 0.1 from the mean.
Then, we can accept the hypothesis H0 with a probability equal to the mean calculated 0.447. As conclusion, this test confirms our hypothesis of the normality distribution.

D'Agostino's K^2 Test

paths from Dataset1

paths	α	P-value	GSNR Normality check
Dusseldorf-> Leipzing	0.050	0.463	Normal
Hamburg-> Munich	0.050	0.424	Normal
Frankfurt->Ulm	0.050	0.502	Normal

Table 5.6: Normality check using D'Agostino's K^2 Test D1

paths from Dataset0

paths	α	P-value	GSNR Normality check
Essen->Bremen	0.050	0.510	Normal
Ulm->Munich	0.050	0.091	Normal
Berlin->Hannover	0.050	0.136	Normal

Table 5.7: Normality check using D'Agostino's K^2 Test D0

The normality testing is done over all our files obtained, for each connection generated by each of dataset1 and dataset0. However, the tables above 5.6 and 5.7 present some randomly chosen paths with their corresponding normal test from the original dataset.

The table 5.8 shows the averaged results of p-value for dataset0 and dataset1. As we expected, this test shows also that our dataset of GSNR in both dataset follow always a Gaussian distribution with a mean of p-value greater than the

Dataset	$\mu_{P-value}$	$\sigma_{P-value}$
D0	0.398	0.11
D1	0.523	0.09

Table 5.8: Normality check using Shapiro-Wilk test

threshold value of 0.05. In this way, we proved the assumption of the CLT that the values of our data sample must accumulate normally when it is big enough with independent variables.

5.3 Comparison between GSNR simulated on full-path from d1 and predicted along sub-paths from d0:

In this experiment we predict analytically the GSNR 4.2.4 in order to check it's compatibility with the GSNR simulated by the system from an experimental point of view by applying Montecarlo along a full path connection taken from dataset1. Hence, for each propagated lightpath, we apply the formulas of regression 4.2 and 4.4 on the data collected from dataset0 to predict the mean and the standard deviation from sub-paths for the whole path of dataset1. For example, to propagate a lightpath from Numberg to Berlin:

1) The shortest calculated path obtained by our algorithm is Numberg, Leipzig then Berlin.

2) We collect the distribution of the simulated $GSNR_{simulated}$ applied by the system in which we focus on the optimum applied power.

3) We compute the predicted $GSNR_{analytical}$ mean and standard deviation by regression on the whole path as the following 4.4:

$$GSNR_{Numberg,Berlin} = \frac{1}{\frac{1}{GSNR_{Numberg,Leipzig}} + \frac{1}{GSNR_{Leipzig,Berlin}}} \tag{5.1}$$

while the predicted Standard deviation is calculated as follows:

$$std_{Numberg,Berlin} = \sqrt{{std_{Numberg,Leipzig}}^2 + {std_{Leipzig,Berlin}}^2} \tag{5.2}$$

In fact, the final standard deviation can be the summation of the node to node sub-paths connections because they are uncorrelated, for each lightpath the variance of the total GSNR is the sum for the variances for each span.

4) Finally, we compare all the corresponding predicted accumulated GSNR from sub-paths with the experimental simulated GSNR for the same path as shown in 5.11 and we show it statistically by the following activities.

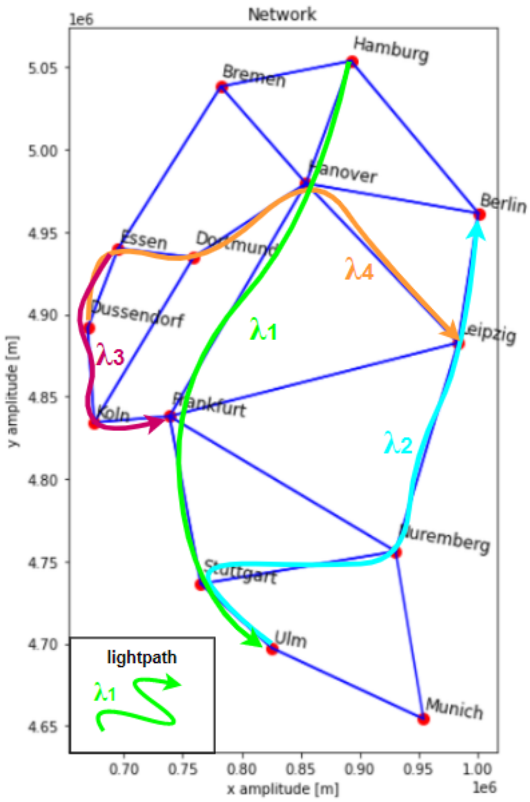

Figure 5.11: Picture of the 4 chosen paths for German

We choose randomly 4 paths from dataset1, and we present the distribution followed by $GSNR_{simulated}$ for each of them against the accumulated $GSNR_{analytical}$ on the sub-paths followed by each lightpath λ from dataset0. The predicted GSNR is normally distributed and is presented by it's mean and standard deviation on top of the $GSNR_{simulated}$ distribution in the following graph 5.11.

5.3.1 Graphical comparison of GSNR distribution

In the following figures, we show the total agreement between the distribution of GSNR coming from dataset1 and the one predicted by regression from dataset0 for the same path connection. Clearly, the red line completely fall on top of the histogram distribution of GSNR in blue.

connection1: λ_1 from Berlin->Ulm

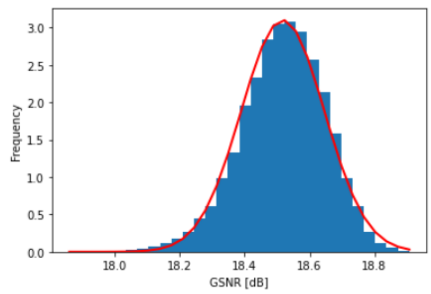

Figure 5.12: Comparison of GSNR distribution for Berlin->Ulm

connection2: λ_2 from Ulm->Hamburg

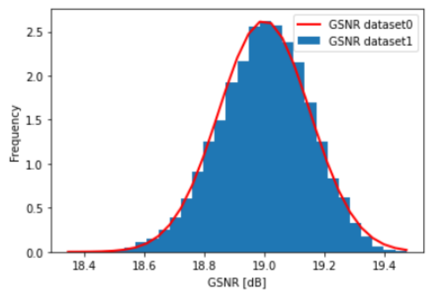

Figure 5.13: Comparison of GSNR distribution for Ulm->Hamburg

connection3: λ_3 from Essen->Frankfurt

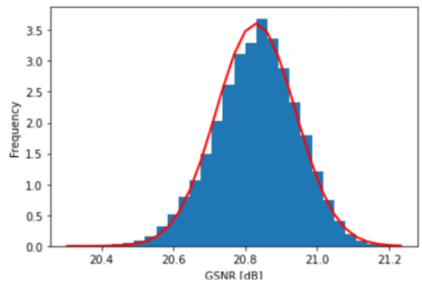

Figure 5.14: Comparison of GSNR distribution for Essen->Frankfurt

connection4: λ_4 from Dusseldorf->Leipzig

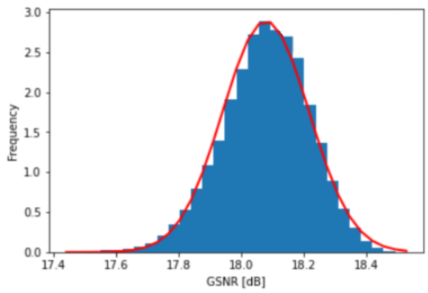

Figure 5.15: Comparison of GSNR distribution for Dusseldorf->Leipzig

According to the graphics above 5.12, 5.13, 5.14 and 5.15 we obviously prove that what we expected statistically converge in a way very similar to the simulated GSNR. The red line shows always the distribution of the statistical computation on top of the distribution of the histogram showing the distribution of the simulated GSNR. This simulated GSNR applied as a QoT for a propagated connection in the routing space is in agreement regarding the normal distribution with what we expect for the nominal GSNR.

5.3.2 Analytical comparison of GSNR distribution: Z-test

Furthermore, in this section we apply a statistical test (Z-test) in order to show the compatibility between the distribution of the estimated GSNR and the simulated GSNR in a more robust way. As shown before, the distribution in any case should be normal, but we seek to show how much these 2 entities have an equivalent matching.

In statistics, the more samples you have, the more reliable the mean is. For that reason, our dataset samples are always formed of 10,000 elements.
We apply the formula mentioned in 4.8, for all the 500 connections taken from dataset1 and compared to each of the calculated GSNR mean and σ over all the sub-paths of the same whole path connection.

We perform the test on the whole dataset1 in which the computations on all the 500 connections output a Z-score that falls in our confidence interval and confirms our hypothesis that the 2 samples are identically shaped, and so they behave as Gaussian following the same distribution.

On top of the analytical identity distribution test, we provide a graphical overview representing the GSNR distributions of the compared samples.

In the following table 5.9, we present 4 randomly chosen lightpath connection paths as presented in figure 5.11 for each λ, in which we show that Z-critical as explained 4.4.1 that should always fall in the interval of confidence $-196 \leq Z_{critical} \leq 1.96$.

lightpath	μ_1	σ_1	μ_2	σ_1	Z-score
λ_1	18.518	0.1283	18.5172	0.1286	0.139
λ_2	18.998	0.151,	18.99846	0.15197	0.068
λ_3	20.831	0.114	20.8257	0.1105	1.056
λ_4	18.078	0.138	18.07138	0.1376	1.074

Table 5.9: Table of the 4 chosen lightpaths for German

As shown by table 5.9, the mean Z-score falls in the interval of confidence that w take on 95% with a mean of 0.297 ± 0.122 ($-196 \leq 0.297 \leq 1.96$) of has a very low average, as the difference of the Gaussians must be low.
The Z-score measurement seems a bit not a stable representation because of the low value of first quartile (also known as the 25_{th} percentile) and correspondingly the high value of the third quartile. This spread is due to the fact we are comparing random chosen paths:
Some of them, taken from dataset 1, are just a line of a couple of city hops, exactly equal to the compared paths from dataset 0. For example from Bremen to Essen, there is no hops to cross and they are directly connected, so the obtained is Z-score= 0.003 very low. Instead, many other lightpaths have to travel a long OLS made of 2 to 6 crossed hops. In this second case, we assist to the analytical GSNR computations that can slightly differ from the experiment ones.

According to the graphical and statistical results 5.9, we conclude that when we have a big dataset for a huge network, instead of calculating the GSNR on every propagated connection, we can simply only compute the distribution on sub-lines and then estimate the full-line GSNR distribution.

5.4 Margin of correction on the analytical GSNR

The observation of Gaussian distribution makes us think of proposing a method for setting the needed margin on a network, given the standard deviation of the GSNR. This margin sets a fixed maximum tolerable out of service percentage.

In our optical network, the network controller estimates a nominal GSNR value. Due to the variation in NEs working point, this estimated nominal GSNR has always some degree of uncertainty. Our predictions must be compared to the real values that the network has as simulated GSNR from dataset1.
This predicted nominal GSNR may be a bit optimistic, so we add a margin of conservation on the estimated nominal GSNR in order to check our system behavior in term of functionalities.

Hence, the GSNR applied by the optical system could be higher or smaller than the nominal predicted $GSNR_{nominal}$. In case our simulated GSNR is smaller, we will face a problem of running out of service. Instead, by applying a GSNR higher than the predicted $GSNR_{nominal}$ we perform lost of our potential resources capacity, but it will not cause a problem regarding the functionality.

The margin of conservation applied is the "Three-sigma" test mentioned in 4.4.2, used to assign a statistical assumption to ensure that the distribution of uncertainties covers all the possibilities for the quality of performance in our optical network topology.
This margin will be calculated by considering the nominal GSNRs predicted in dataset0, then we apply a subtraction of "three-sigma" ($3\sigma_{GSNR}$) for each single connection, in order to set the minimum control limits of 1% in statistical domain for the simulated GSNR of running out of service.
Accordingly, the average over all the applied margin for all the connections is of 0.258 dB with a minimum of 0.147 dB and as maximum 0.28 dB.
Thanks to this limit of conservation, we can calculate the percentage for which our system may go out of service among all the propagated connections.

$$GSNR_{optimal} = GSNR_{nominal} - 3\sigma_{GSNR} \tag{5.3}$$

Accordingly, we collect for each single propagated connection in German topology, among all the generated GSNRs by the Montecarlo runs, the number of connections that has a simulated GSNR higher than the optimal conservative GSNR.

$$\Delta GSNR = GSNR_{actual} - GSNR_{predicted} \tag{5.4}$$

We assign for the number of connections that are successfully propagated the name $N_{Successful}$ all the cases when $\Delta GSNR > 0$.

Hence, we collected the averaged mean μ and the standard deviation σ for the $N_{Successful}$ over all the connections propagated in which they are respectively:

μ= 9963.766 and σ= 7.171.

Hence, this means that in average we have 99.637% of connections are successful among our generated GSNRs from the 10,000 runs in which they exceed the optimal conservative margin that we applied.
Instead, 0.363% of these generated GSNRs are less than the optimal conservative GSNR, which lead for an out of service because the optical system needed more resources to propagate correctly the lightpath.
We conclude that in general the probability of running out of service with a margin = 3σ is p_{oos}=0.363%.

Therefore, here we present an example of a random connection going from 'Numberg' to 'Essen' 4.2, taken from German DT topology:

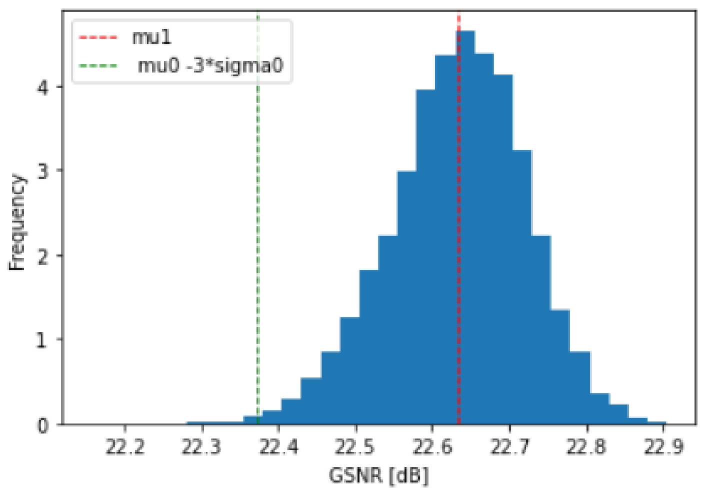

Figure 5.16: GSNR simulated from dataset1 from Numberg to Essen

Thus, the figure above shows operative conservative GSNR with margin $GSNR_{operative}$ = 22.373 in green on top of the distribution obtained for the GSNR

generated for this specific connection with mean shown by the red line of μ_1= 22.533.
Also, this connection has a $N_{Successful}$= 9975 (99.75%), so clearly the tail of the GSNR Gaussian distribution shows the maximum allowable oos= 0.25% of unaccepted propagated lightpaths that goes out of service from all the 10,000 random generated.

After all, this margin is applied in order to correct and to enhance the prediction of GSNR for our propagated connections in the routing space.
Hence, we notice graphically that more than 99% of the cases are above our observative extreme margin of conservation 5.16 and less than 1% (the tail) of the generated GSNRs are smaller.

5.5 Euclidean distance of the GSNRs between dataset1 and dataset0:

In order to have a metric of our calculation of the conservative margin of the GSNR, we apply the Euclidean distance between the optimal and the simulated GSNRs.

This metric is applied to check the error that we make if we use the conservative GSNR for all the dataset1. The error is the distance between our parameters in which is represented in term of RMSE as mentioned in 4.4.2 and is done for all the generated connections in dataset1.

In general, the higher is the value of the RMSE the poorer is the ability of the system to predict correctly the analytical GSNR from dataset0.
Anyway, we collect the mean and standard deviation on all the 500 collected values of RMSE and what we obtain is a: μ_{RMSE}= 0.3004 dB and σ_{RMSE}= 0.0976.

Clearly, what we obtain is a small value of the μ_{RMSE}= 0.3 dB with a deviation around the mean of 0.097 which reflects that the model can relatively predict the data accurately because the closer is the value of RMSE to 0 the more accurate is the model in predicting results.
It seems that our system is able to predict very accurate results, and as consequence, we are able to propagate the lightpath successfully from a source ROADM to the destination without going out of service more than 0.37% as was proved by the applied conservative margin.
So, the error between the observative GSNR and the applied GSNR over all the 10,000 GSNRs realization is very little shown by our accurate results.
We present as an example, the connection from 'Koln' to 'Ulm' in German topology 4.2. For this propagated lightpath, the RMSE between the GSNR predicted by our system and the simulated GSNRs realization is RMSE= 0.27 dB. That obviously

means that what we expected as optimal conservative GSNR to propagate this lightpath, is accurate with the simulated propagated GSNR over the routing space. Hence, we conclude that the connection can be successfully propagated starting from our prediction with a small RMSE .

5.6 Transfer Learning results for US topology

5.6.1 Machine learning accuracy

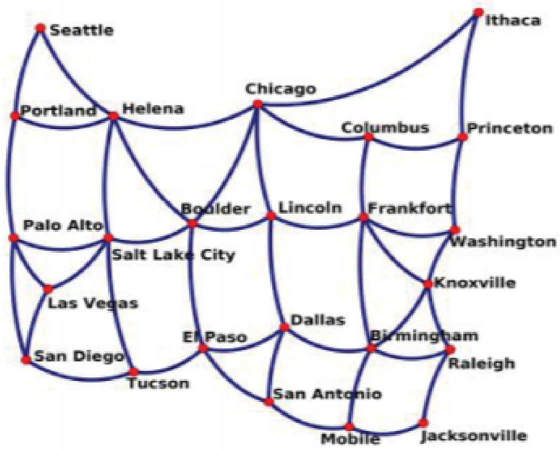

Figure 5.17: US optical network topology

The mainly aim of our ML application is to effectively train a known model to make it able to have a reliable QoT-E in order to correct the GSNR uncertainties in the network controller for an unknown network. However, As we have a dataset on the German topology network, we exploit these data to do a transfer learning on a new network topology US network to estimate the GSNR.

We generate he training dataset from German topology and is done by 500 connections. Hence, in order to have the same generated input connector losses and ripples on every shared line, we consider them fixed without including uncertainties. Similarly, we create a testing dataset from US network 5.17, and as done for the training dataset, we fix both input connector losses and gain ripples. For each of these datasets, we compute the QoT GSNR and then we feed our machine learning model with these data to be able to estimate the QoT for US network topology.

Example from training dataset: German topology

N_{fibers}	$GSNR$ [dB]	Ripples [dB]	Losses [dB]
7	20.14	0.987	0.749
5	21.644	0.987	0.749

Table 5.10: Examples of training dataset for connections in German

Example from testing dataset: US topology

N_{fibers}	$GSNR$ [dB]	Ripples [dB]	Losses [dB]
45	9.685	0.987	0.749
41	10.530	0.993	0.75

Table 5.11: Examples of testing dataset for connections in US network

The tables above present an example of the trained and tested dataset respectively 5.10 and 5.11 for some connections after collecting the feature (GSNR) and the labels (number of fiber spans, constant gain ripple and input connect loss).

The used DNN is characterized by several optimized parameters, considering as training steps = 1000 and a default learning rate = 0.01.
Although, the number of hidden layers in DNN is a very essential characteristic. After trying multiple models with different number of hidden layers and neurons we arrive to a trade-off between the performance of our algorithm and the time needed to execute all the predictions. Hence, an increase in the number of these parameters may overfit the network, especially if the network dataset is not very big as in our case 500 training and 100 testing. So, we decide for 3 hidden layers each with 3 neurons as an optimum option for our dataset.

Firstly, we divide in three subsets of the German network dataset, then we perform the training, validation, and testing; we apply as 70% training ratio and 30% testing ratios as proportions.

Accordingly, we collect for each single propagated connection in USnetwork topology the number of connections from the testing dataset that has a GSNR higher than the $GSNR_{predicted}$ by ML algorithm applying the following formula:

$$\Delta GSNR = GSNR_{actual} - GSNR_{predicted} \tag{5.5}$$

in which $GSNR_{actual}$ is the GSNR calculated for a lightpath along the OLS with fixed gain ripples and input connector losses. We assign for the number of connections that are successfully propagated the name $N_{Successful}$ all the cases when $\Delta GSNR > 0$.
What we obtain is a $\Delta GSNR$=89% with an accuracy in term of RMSE=0.54 dB. This means that in average, 11% of the propagated connections may run oos.
The penalty obtained when running oos by applying the transfer learning from a known network to another new network, opens a path to apply a margin of conservation for a more conservative architecture with more accurate predictions.

5.6.2 Statistical analysis on USnetwork

In this section, we consider the new topology "US network" that we base our study on. We consider the input connector losses and the gain ripples as variables with uncertainties. We collect a dataset for US network over all the connections propagated, by computing the GSNR over a full line from ROADM-to-ROADM connection and we call the GSNR as $GSNR_{actual}$. Then, we do the statistical analytical analysis in which we create another dataset as done for dataset0 by considering the GSNR of sub-paths, node-to-node connections 5.1.2.
On top of the estimated GSNR in dataset0, we apply the margin of conservation $3\sigma_{GSNR}$ in order to check the extreme conservative GSNR and to put a maximum of going out of service for our propagated connections using the formula:

$$\Delta GSNR_{conservative} = GSNR_{predicted} - 3\sigma_{GSNR} \tag{5.6}$$

In order to check the degree of conservation that we obtain from the statistical approach, we compute the difference between the $GSNR_{actual}$ already computed for each connection considering fixed losses and ripples and the calculated $GSNR_{conservative}$. This $GSNR_{actual}$ is used also for the comparison with the ML predicted GSNR.

$$\Delta GSNR = GSNR_{actual} - GSNR_{conservative} \tag{5.7}$$

as done before, we check the cases where $\Delta GSNR$>0 and we consider such connection as successful avoiding running out of service with a margin of conservation. Hence, we obtain 99% of the connections that propagate successfully with a $\Delta GSNR$>0. We also compute the RMSE to check the error difference between the GSNRs, and we obtain a RMSE= 0.37 dB.

We compare the percentage of difference obtained by the transfer learning to what we obtain statistically. Hence, what we predict by the statistical method with a margin of conservation seems to be more accurate and more consistent with the

$GSNR_{actual}$ with constant gain ripples and input connector loss. Moreover, the error RMSE= 0.58 dB committed by the ML results is higher than the accuracy RMSE= 0.37 dB done by applying the statistical approach.
So, the ability of the transfer learning to predict our QoT GSNR seems to be weaker than applying the statistical approach along each OLS lightpath propagated with a conservative margin.

Chapter 6

Conclusions

In this thesis we investigated the problem of uncertainties of the connector input losses and the ripples introduced respectively by the fibers and amplifiers, in a real network scenario. We assigned the GSNR as a QoT for lightpaths propagated over the routing space in real network scenarios.

6.1 GSNR computation by Simulated method

We started by generating a dataset of 500 random connections from the German topology, applying the shortest path algorithm. On top of the generated connections.
For this purpose, we applied a Montecarlo method with a high number of runs to obtain a distribution of our data, assuming Gaussian fluctuations around reference values for both input connector loss and ripples aiming for analysing the GSNR distribution. We inspected separately the ASE and NLI noises introduced during the signal propagation and considering always the optimal power in which we have a maximum GSNR and by considering only the cut-channel as a channel under test.

Additionally, we devised and implemented a test for Gaussianity for the GSNR metric, based on the Characteristic Function cf, which interprets the problem in a more robust way respect to the graphical methods proposed. Additionally, we have shown this test to be a very powerful check for Gaussianity.
Hence, by using two graphical shows (histogram and Q-Q) and two statistical methods (Shapiron-Wilk and Agostino k^2 test) we proved that for each lightpath connection, the GSNR generated follows a perfectly normal distribution according to the CLT.

6.2 GSNR computation by Statistical method

In general, in case of a huge topology with a big number of connections propagated, it is very expensive in terms of computational complexity to compute the GSNR. For this motivation, we created a second estimated dataset for the same connections, with the difference that here we applied a simulation along the path between each node-to-node connection.
What we care about was to ensure that the Monte Carlo approach is directly consistent with the implemented statistical analytical model. Hence, between the simulated GSNR and the estimated nominal GSNR.
For this reason, we compared the distribution of GSNR estimated and simulated graphically by the bell-curve and by using the Z-test model with a 95% as confidence interval.
Graphically, we obtained a perfect alignment by plotting the distribution of the predicted GSNR using the mean and standard deviation on top of the generated GSNR distribution simulated. Moreover, we acquired from the Z-test a z-score that always falls in the interval of confidence with a mean of 0.3 and a standard deviation of 0.2. Hence, this means that what we estimated was very consistent with what the network simulates.

Moreover, the estimated nominal GSNR could be a bit optimistic, so we added on top of it a margin of conservation to check the functionality and to save our system behavior.
Hence, this minimum margin deployed to decrease the uncertainty introduced in the GSNR computation for each LP, and so to enable a reliable path computation to rely on the overall LP GSNR.
Therefore, to deploy more traffic by an optimum exploitation of the installed equipment, we must apply a lower system margin. The margin of conservation is applied on all the propagated LP by decrementing the 3σ limit for each estimated GSNR to obtain an optimal working point for each connection.
The average applied margin is of 0.258 with a minimum of 0.147 dB and as maximum 0.281 dB.
For all the connections, we obtained in average p_{oos}=0.363% and 99.637% of connections are successful among our generated dataset GSNRs that they exceed the optimal GSNR after the application of the conservative margin.

Aiming for having a metric for the distance between the optimal and the simulated GSNR, we applied the Euclidean distance between to check the error that we make if we use the conservative GSNR for all the dataset1. The error is the distance between our parameters and is represented in term of RMSE on top of all the propagated connections over the routing space.

What we achieved is μ_{RMSE}= 0.3004 dB and σ_{RMSE}= 0.0976. In fact, the lower is the value of the RMSE, the better is the ability of the system to predict correctly the analytical GSNR from this dataset.
Clearly, our system is able to predict very accurate results, and as consequence, we are able to propagate the lightpath successfully from a source ROADM to the destination without going out of service more than 0.37% as was proved by the applied conservative margin on top of the nominal estimated GSNR.

6.3 Transfer learning and statistical correction on USnetwork

In this work, we proposed the application of a transfer learning from a trained dataset comes from German topology to correct the GSNR computation by the QoT-E in an unknown network topology USnetwork. We generated dataset synthetically for both networks, for training and testing purposes.
For both dataset we have fixed all the uncertainties coming from input connector loss and gain ripples in order to guarantee the same values for the shared sub-paths on different propagated connections.

The results obtained by the transfer learning shows the low capability of our DNN model of predicting the GSNR with a high penalty of going out of service.
We extend our study to apply the statistical method with a margin of conservation in order to obtain a QoT-E. The predicted GSNRs are collected by considering the uncertainty on the input connector losses and the gain ripples, then compared to the same actual GSNR we used to check the system behavior in the ML model.
Accordingly, the statistical method is more accurate and more reliable in predicting the QoT and by reducing the uncertainty done by the fluctuations coming from the NEs in our optical network and it gives way lower p_{oos} and RMSE.

6.4 Future work, ways of improving

What we mainly proved in our work is that the GSNR follows a Gaussian distribution independently line-by-line system by considering the uncertainties on the connector input loss and the ripples. In general, when we can rely on a full knowledge of the physical layer and it is statistics, it opens a green field to exploit this additional knowledge in order to minimize the margin applied on the network.

A way of improving our study can be achieved by doing a full statistical regression beside finding the statistics of the GSNR that accumulates on the network that we have done and by checking the statistic of the uncertainty of some components.

Nowadays, typically in optical networks, when we know the nominal GSNR we should apply a margin on top of it, because we considered the fluctuations within the network and we apply a GSNR smaller than the nominal one. This smaller GSNR introduces a huge loss of capacity because we have to stay conservative.
Instead, if we can have an exact characterisation of the statistics and we can have a full knowledge of the fluctuations that affects GSNR on a lightpath, we expect an exact prediction of GSNR. In general, the optical network infrastructure is already installed but with high traffic demand. The operators must keep the infrastructure trying to have a return of the money spent on it.

We can come up with a better way to exploit all the network capacity from the infrastructure having already deployed the physical layer.
A good strategy for decreasing the uncertainty introduced by the fluctuations, using GnPy model as a prediction of the QoT and so we can work on improving the QoT-E and then on filling this gap of knowledge which is a big objective.
This approach takes data from the field to reduce the uncertainty. This means larger GSNR, larger bit rate, then deploy larger traffic in a reliable way and then gain more money.

Moreover, for future analyses the considered set of configurations can be further enhanced and enlarged by introducing Raman amplification alongside EDFAs and by considering different fiber types and lengths taking into account the inter channel stimulated Raman scattering.

Finally, in order to make the studies wider, the uncertainties introduced by the input connector losses and the gain ripples can be extended by including any kind of PDF not only normal.

Bibliography

[1] «Fiber-Optic Technologies». In: URL: `https://www.ciscopress.com/articles/article.asp?p=170740&seqNum=2` (cit. on p. 6).

[2] V. Curri. In: *Open optical network - course lessons*. 2021 (cit. on pp. 7, 10, 13, 16, 17, 21, 23, 25).

[3] A. D'Amico et al. «Enhancing Lightpath QoT Computation with Machine Learning in Partially Disaggregated Optical Networks». In: *IEEE Open Journal of the Communications Society* (2021), pp. 1–1. DOI: `10.1109/OJCOMS.2021.3066913` (cit. on pp. 7, 17, 33, 35, 48).

[4] Chava Vijaya Saradhi and Suresh Subramaniam. «Physical layer impairment aware routing (PLIAR) in WDM optical networks: issues and challenges». In: *IEEE Communications Surveys Tutorials* 11.4 (2009), pp. 109–130. DOI: `10.1109/SURV.2009.090407` (cit. on pp. 8, 10, 14, 15, 17–19).

[5] University of Windor. «Optical Networks – Basic Concepts (Part 1)». In: pp. 1–49. URL: `https://dokumen.tips/documents/optical-networks-basic-concepts-part-1-download-movie-optical-fibers.html` (cit. on p. 10).

[6] Vittorio Curri. «Software-Defined WDM Optical Transport in Disaggregated Open Optical Networks». In: *2020 22nd International Conference on Transparent Optical Networks (ICTON)*. 2020, pp. 1–4. DOI: `10.1109/ICTON51198.2020.9203450` (cit. on p. 12).

[7] Marc De Leenheer, Yuta Higuchi, and Guru Parulkar. «An open controller for the disaggregated optical network». In: *2018 International Conference on Optical Network Design and Modeling (ONDM)*. 2018, pp. 230–233. DOI: `10.23919/ONDM.2018.8396136` (cit. on p. 12).

[8] Sami Mumtaz, René-Jean Essiambre, and Govind P. Agrawal. «Nonlinear Propagation in Multimode and Multicore Fibers: Generalization of the Manakov Equations». In: *Journal of Lightwave Technology* 31.3 (2013), pp. 398–406. DOI: `10.1109/JLT.2012.2231401` (cit. on p. 16).

[9] «non linear effects». In: URL: `https://doi.org/10.1007/3-540-46629-0_9` (cit. on p. 19).

[10] «Weighted graphs». In: URL: `https://courses.cs.vt.edu/~cs3114/Fall10/Notes/T22.WeightedGraphs.pdf` (cit. on p. 27).

[11] Gregory Cowle. «The state of the art of modern non-SDM amplification technology in agile optical networks: EDFA and Raman amplifiers and circuit packs». In: *2017 Optical Fiber Communications Conference and Exhibition (OFC)*. 2017, pp. 1–28 (cit. on p. 29).

[12] Danshi Wang, Zhiguo Zhang, Min Zhang, Meixia Fu, Jin Li, Shanyong Cai, Chunyu Zhang, and Xue Chen. «The Role of Digital Twin in Optical Communication: Fault Management, Hardware Configuration, and Transmission Simulation». In: *IEEE Communications Magazine* 59.1 (2021), pp. 133–139. DOI: `10.1109/MCOM.001.2000727` (cit. on p. 30).

[13] «Overview of telecom infra project». In: URL: `https://telecominfraproject.com/` (cit. on p. 31).

[14] «telecom infra project». In: URL: `https://telecominfraproject.com/` (cit. on p. 31).

[15] «telecom infra project». In: URL: `https://github.com/Telecominfraproject/gnpy` (cit. on p. 31).

[16] Jean-Luc Auge, Gert Grammel, Esther le Rouzic, Vittorio Curri, Gabriele Galimberti, and James Powell. «Open Optical Network Planning Demonstration». In: *2019 Optical Fiber Communications Conference and Exhibition (OFC)*. 2019, pp. 1–3 (cit. on p. 32).

[17] P. Poggiolini, G. Bosco, A. Carena, R. Cigliutti, V. Curri, F. Forghieri, R. Pastorelli, and S. Piciaccia. «The LOGON strategy for low-complexity control plane implementation in new-generation flexible networks». In: *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*. 2013, pp. 1–3. DOI: `10.1364/OFC.2013.OW1H.3` (cit. on p. 32).

[18] J. Strand, A.L. Chiu, and R. Tkach. «Issues for routing in the optical layer». In: *IEEE Communications Magazine* 39.2 (2001), pp. 81–87. DOI: `10.1109/35.900635` (cit. on p. 32).

[19] Emilio Riccardi, Paul Gunning, Óscar González de Dios, Marco Quagliotti, Víctor López, and Andrew Lord. «An Operator view on the Introduction of White Boxes into Optical Networks». In: *Journal of Lightwave Technology* 36.15 (2018), pp. 3062–3072. DOI: `10.1109/JLT.2018.2815266` (cit. on p. 33).

[20] C. Manso, R. Munoz, N. Yoshikane, R. Casellas, R. Vilalta, R. Martinez, T. Tsuritani, and I. Morita. «TAPI-enabled SDN control for partially disaggregated multi-domain (OLS) and multi-layer (WDM over SDM) optical networks». In: *IEEE/OSA Journal of Optical Communications and Networking* 13.1 (2021), A21–A33. DOI: `10.1364/JOCN.402187` (cit. on p. 33).

[21] P. Poggiolini, G. Bosco, A. Carena, R. Cigliutti, V. Curri, F. Forghieri, R. Pastorelli, and S. Piciaccia. «The LOGON strategy for low-complexity control plane implementation in new-generation flexible networks». In: *2013 Optical Fiber Communication Conference and Exposition and the National Fiber Optic Engineers Conference (OFC/NFOEC)*. 2013, pp. 1–3. DOI: `10.1364/OFC.2013.OW1H.3` (cit. on p. 39).

[22] «Central limit theorem». In: URL: `https://www.sciencedirect.com/topics/engineering/central-limit-theorem` (cit. on p. 42).

[23] P. Mach and H. Hochlova. «Testing of normality of data files for application of SPC tools». In: *27th International Spring Seminar on Electronics Technology: Meeting the Challenges of Electronics Technology Progress, 2004.* Vol. 2. 2004, 318–321 vol.2. DOI: `10.1109/ISSE.2004.1490443` (cit. on p. 43).

[24] Mohammad Ebrahim Hajiabadi and Habib Rajabi Mashhadi. «Analysis of the Probability Distribution of LMP by Central Limit Theorem». In: *IEEE Transactions on Power Systems* 28.3 (2013), pp. 2862–2871. DOI: `10.1109/TPWRS.2013.2252372` (cit. on pp. 43, 52).

[25] M.J. Arnold, D.R. Iskander, and A.M. Zoubir. «Testing Gaussianity with the characteristic function». In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. 1995, 2012–2015 vol.3. DOI: `10.1109/ICASSP.1995.480670` (cit. on p. 43).

[26] Ralph D'Agostino and E. S. Pearson. «Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1». In: vol. 60. 3. [Oxford University Press, Biometrika Trust], 1973, pp. 613–622 (cit. on p. 44).

[27] Ralph B. D'Agostino. «An Omnibus Test of Normality for Moderate and Large Size Samples». In: vol. 58. 2. [Oxford University Press, Biometrika Trust], 1971, pp. 341–348 (cit. on p. 44).

[28] «Regression analysis». In: URL: `https://corporatefinanceinstitute.com/resources/knowledge/finance/regression-analysis/` (cit. on p. 45).

[29] «k-test». In: URL: `http://homework.uoregon.edu/pub/class/es202/ztest.html` (cit. on p. 45).

[30] «Understanding Z-Test». In: URL: `https://www.investopedia.com/terms/z/z-test.asp` (cit. on p. 46).

[31] «three-sigma-limits». In: URL: `https://www.investopedia.com/terms/t/three-sigma-limits.asp` (cit. on p. 46).

[32] Rubing Huang, Chenhui Cui, Weifeng Sun, and Dave Towey. «Poster: Is Euclidean Distance the best Distance Measurement for Adaptive Random Testing?» In: *2020 IEEE 13th International Conference on Software Testing, Validation and Verification (ICST)*. 2020, pp. 406–409. DOI: `10.1109/ICST46399.2020.00049` (cit. on p. 48).

[33] «Q-Q plot distribution». In: URL: `https://data.library.virginia.edu/understanding-q-q-plots/` (cit. on p. 55).