FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

ANDRÉ LUCAS PINHO FERNANDES

Designing Feasible Deployment Strategies for Cell-Free Massive MIMO Networks: Assessing Cost-Effectiveness and Reliability

TD 12/2024

UFPA / ITEC / PPGEE
Guamá University Campus
Belém-Pará-Brasil
2024

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"PROJETANDO ESTRATÉGIAS DE IMPLANTAÇÃO VIÁVEIS PARA REDES MIMO MASSIVAS SEM CÉLULAS: AVALIANDO A RELAÇÃO CUSTO-BENEFÍCIO E A CONFIABILIDADE"

AUTOR: **ANDRÉ LUCAS PINHO FERNANDES**

TESE DE DOUTORADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE TELECOMUNICAÇÕES.

APROVADA EM: 14/06/2024

TD 12/2024

BANCA EXAMINADORA:

DocuSigned by: João Weyl
EFE1BE0196F444C...

Prof. Dr. João Crisóstomo Weyl Albuquerque Costa
(Orientador – PPGEE/UFPA)

DocuSigned by: Paolo Monti
A2CACAF683AA415...

Prof. Dr. Paolo Monti
(Coorientador - CHALMERS UNIVERSITY OF TECHNOLOGIES/SUÉCIA)

DocuSigned by: Jasmine Araújo
CEDA51E483D843B...

Prof.ª Dr.ª Jasmine Priscyla Leite de Araújo
(Avaliadora Interna - PPGEE/UFPA)

DocuSigned by: Fabrício Farias
32606AA841D646A...

Prof. Dr. Fabrício de Souza Farias
(Avaliador Externo ao Programa - CAMPUS CAMETÁ/UFPA)

DocuSigned by: Roberto Menezes Rodrigues
CF90C6E0833449D...

Prof. Dr. Roberto Menezes Rodrigues
(Avaliador Externo ao Programa - FEEB/UFPA)

DocuSigned by: André Mendes Cavalcante
E7CD59D948CE4D9...

Dr. André Mendes Cavalcante
(Avaliador Externo - ERICSSON)

DocuSigned by: Daniel Benevides da Costa
6F84E64849C6412...

Prof. Dr. Daniel Benevides da Costa
(Avaliador Externo - KING FAHD UNIVERSITY OF PETROLEUM & MINERALS)

VISTO:

DocuSigned by: Diego Lisboa Cardoso
1C8C1AE22FB2444...

Prof. Dr. Diego Lisboa Cardoso
(Coordenador do PPGEE/ITEC/UFPA)

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

F363d Fernandes, André Lucas Pinho.
Designing feasible deployment strategies for cell-free massive MIMO networks : Assessing cost-effectiveness and reliability / André Lucas Pinho Fernandes. — 2024.
XLIII, 170 f. : il. color.

Orientador(a): Prof. Dr. João Crisóstomo Weyl Alburque Costa
Coorientador(a): Prof. Dr. Paolo Monti
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2024.

1. Cell-free massive MIMO. 2. Network deployment. 3. Protection schemes. 4. Reliability assessment. 5. Techno-economic assessment. I. Título.

CDD 621.3811

FEDERAL UNIVERSITY OF PARÁ
INSTITUTE OF TECHNOLOGY
POSTGRADUATE PROGRAM IN ELECTRICAL ENGINEERING

ANDRÉ LUCAS PINHO FERNANDES

Designing Feasible Deployment Strategies for Cell-Free Massive MIMO Networks: Assessing Cost-Effectiveness and Reliability

Doctoral thesis submitted to the Examining Committee of the Postgraduate Program in Electrical Engineering at the Federal University of Pará to obtain the Ph.D. Degree in Electrical Engineering. Area of concentration: Telecommunications.

TD 12/2024

Advisor: João Crisóstomo Weyl Albuquerque Costa

UFPA / ITEC / PPGEE
Guamá University Campus
Belém-Pará-Brasil
2024

This work is dedicated to God, my family, and all those who believed on its development

Acknowledgements

Throughout my life as a researcher, I have been fortunate to receive abundant support and numerous opportunities, for which I am deeply grateful.

First and foremost, I thank God.

I thank my parents, sisters, and grandparents for their continuous encouragement and support.

I thank my wife, Aline Ohashi, for her endless support and dedication throughout this journey.

I thank Professor João Weyl for guiding me in my research and opening up many doors along the way.

I thank all the friends I made throughout my journey, especially Daynara Souza and Marx Miguel, with whom I have not only collaborated on numerous published works but also shared countless laughs during our times away from work.

I am thankful to Professor Paolo Monti for hosting me during my sandwich period abroad and for his significant contributions to my research.

I thank the High-Performance Computer Center of the Federal University of Pará for granting me access to computational resources to develop my work.

Lastly, I appreciate the financial support provided by CNPq, CAPES, Fapespa, Ericsson, and UFPA, which has been essential in my academic pursuits.

“ My entire life consisted of musings, calculations, practical works and trials. Many questions remain unanswered, many works are incomplete or unpublished. The most important things still lie ahead.”

(Konstantin Tsiolkovsky)

Resumo

Redes *cell-free Massive Multiple-Input Multiple-Outputs* (mMIMO) são uma solução promissora para os sistemas móveis de Sexta Geração (6G) e além. Estas redes utilizam múltiplas antenas distribuídas para transmitir e receber sinais de forma coerente, sob um paradigma de comunicação aparentemente não celular que elimina o conceito tradicional de células em redes móveis. Esta mudança apresenta desafios significativos de implementação, pois as ferramentas convencionais projetadas para sistemas celulares são inadequadas para planejar e avaliar arquiteturas *cell-free* mMIMO. Nesse sentido, a literatura vem desenvolvendo modelos específicos para sistemas *cell-free* mMIMO que lidam com a coordenação de sistema, a sinalização fronthaul, as complexidades computacionais necessárias de procedimentos de processamento, o fronthaul segmentado, a transição de implementação a partir de redes celulares e a integração com a tecnologia *Open Radio Access Network* (O-RAN). Esses avanços são fundamentais para transformar o *cell-free* mMIMO de um sistema teórico em uma aplicação prática. Apesar disso, mais estudos são necessários para integrar os modelos existentes e desenvolver ferramentas práticas de avaliação para examinar a viabilidade de *cell-free* mMIMO e de seus facilitadores. Esta tese visa preencher estas lacunas ao propor novas ferramentas para avaliar a viabilidade de redes *cell-free* mMIMO em termos de confiabilidade e custos.

A primeira ferramenta concentra-se na avaliação da confiabilidade do *cell-free* mMIMO. Ela é usada para melhorar a compreensão de possíveis impactos de falhas e para desenvolver esquemas de proteção eficazes para o fronthaul de redes *cell-free* mMIMO. Resultados para uma implementação em escritório *indoor* com uma área de 100 m^2 e espaçamento entre *Transmission-Reception Points* (TRPs) de 20 m, demonstram que sistemas *cell-free* com fronthaul segmentado, ou seja, com conexões de fronthaul seriais entre TRPs, necessitam de estratégias de proteção. É mostrado que interconectar cadeias seriais e duplicar parcialmente as cadeias seriais (redundância de 40%) são esquemas de proteção eficazes. Por fim, nos cenários *indoor* considerados, a interconexção parece ser a alternativa mais viável quando o número de cadeias seriais é maior que três.

A segunda ferramenta avalia o *Total Cost of Ownership* (TCO) de redes *cell-free* mMIMO, considerando aspectos essenciais, como demandas dos usuários, limitações de largura de banda de fronthaul e capacidades de processamento de *hardware*. A ferramenta é usada para avaliar os custos de duas divisões funcionais da literatura que são equivalentes a arquiteturas de processamento distribuído e centralizado para redes *cell-free* mMIMO. Resultados para um cenário urbano ultradenso cobrindo uma área de 0,25 km^2 com até 800 TRPs, revelam que processamento centralizado é mais viável para a maioria das demandas dos usuários, configurações de *hardware* de TRP, e considerações de custo. Apesar disso, processamento distribuído pode ser mais viável em casos limitados de baixa demanda (até 50 Mbps por usuário) e sob enormes reduções de custos relacionadas à implantação dos TRPs.

Keywords: *Cell-free massive* MIMO, Divisões funcionais, Implantação de rede, Esquemas de proteção, Avaliação de confiabilidade, Fronthaul segmentado, Avaliação técnico-econômica

Abstract

Cell-free Massive Multiple-Input Multiple-Output (mMIMO) networks are a promising solution for the Sixth Generation of mobile systems (6G) and beyond. These networks utilize multiple distributed antennas to transmit and receive signals coherently, under an apparently non-cellular communication paradigm that eliminates the traditional concept of cells in mobile networks. This shift poses significant deployment challenges, as conventional tools designed for cellular systems are inadequate for planning and evaluating cell-free mMIMO architectures. In this sense, the literature has been developing models specific to cell-free mMIMO that deal with system coordination, fronthaul signaling, required computational complexities of processing procedures, segmented fronthaul, transitioning from cellular network deployments, and integration to Open Radio Access Network (O-RAN) technologies. These advancements are instrumental in transforming cell-free mMIMO from a theoretical system to a practical application. Despite this, further study is needed to integrate existing models and develop practical evaluation tools to assess the feasibility of cell-free mMIMO and its enablers. This thesis addresses these gaps by proposing new tools to evaluate the feasibility of cell-free mMIMO networks regarding reliability and costs.

The first tool focuses on evaluating the reliability of cell-free mMIMO. It is used to improve the understanding of possible failure impacts and to develop effective protection schemes for the fronthaul network of cell-free mMIMO networks. Results for an indoor office implementation with an area of 100 m^2 and a Transmission-Reception Point (TRP) spacing of 20 m, demonstrate that cell-free systems with segmented fronthaul, i.e., with serial fronthaul connections between TRPs, require protection strategies. It is shown that interconnecting serial chains and partially duplicating serial chains (40% redundancy) are effective protection schemes. Finally, in the considered indoor scenarios, interconnection appears to be the most feasible alternative when the number of serial chains is higher than three.

The second tool assesses the Total Cost of Ownership (TCO) of cell-free mMIMO and its enablers, considering essential aspects, like user demands, fronthaul bandwidth limitations, and hardware processing capacities. The tool is used to evaluate the costs of two functional splits from the literature that are equivalent to distributed and centralized processing architectures for cell-free mMIMO networks. Results for an ultra-dense urban scenario covering an area of 0.25 km^2 with up to 800 TRPs, reveal that centralized processing is more feasible for most user demands, hardware configurations of TRP, and cost considerations. Despite this, distributed processing may be more feasible in limited cases of low demand (up to 50 Mbps per user) and under massive cost reductions for expenses related to TRPs deployment.

Keywords: Cell-free massive MIMO, Functional splits, Network Deployment, Protection Schemes, Reliability Assessment, Segmented fronthaul, Techno-economic Assessment

Contents

List of figures

List of tables

List of acronyms

1G First Generation of Mobile Systems

2G Second Generation of Mobile Systems

3G Third Generation of Mobile Systems

3GPP Third Generation Partnership Project

4G Fourth Generation of Mobile Systems

5G Fifth Generation of Mobile Systems

AR Augmented Reality

ASD Angular Standard Deviation

ADC Analog-to-Digital Converter

ASD Angular Standard Deviation

AQNM Additive Quantization Noise Model

AWGN Additive White Gaussian Noise

BBU Baseband Unit

BCPU Baseband on Central Processing Unit

BFS Breadth-First Search

BS Base Station

BSV Breakeven Subscription Value

BTRP Baseband on Transmission-Reception Point

CA Carrier Aggregation

CAPEX Capital Expenditures

CC Cross-Connection

CDF Cumulative Distribution Function

CO Central Office

CoMP Coordinated MultiPoint

CoMP-JT Coordinated MultiPoint-Joint Transmission

CP Cyclic Prefix

CPU Central Processing Unit

C-RAN Centralized Radio Access Network

CSI Channel State Information

CSIT Channel State Information at the Transmitter

CTMC Continuous-Time Markov Chain

CU Cost Unit

CeUn Central Unit

DCC Dynamic Cooperation Clustering

DFS Depth-First Search

DSP Digital Signal Processors

D-RAN Distributed Radio Access Network

DAC Digital-to-Analog Converter

DU Distributed Unit

DFT Discrete Fourier Transform

DTMC Discrete-Time Markov Chain

eCPRI Enhanced Common Public Radio Interface

ERSS Ericsson Radio Stripe System

FD Full Duplication

FDD Frequency Division Duplex

FEC Forward Error Correction

FFT Fast Fourier Transform

FTTH Fiber to the Home

FTTB Fiber to the Building

GCC Global Cloud Controller

GPP General Purpose Processor

GOPS Giga Operations Per Second

HARQ Hybrid Automatic Repeat Request

HetNet Heterogenous Network

IMT International Mobile Telecommunications

InH-open Indoor Hotspot Open Office

IFFT Inverse Fast Fourier Transform

IP Internet Protocol

IQ In-Phase and Quadrature

IRR Internal Rate of Return

ISI Inter-Symbol Interference

IT Information Technology

ITU-R International Telecommunication Union - Radiocommunication Sector

I/O Input/Output

kWh Kilowatt Hour

L-MMSE Local-Minimum Mean Square Error

LoS Line-of-Sight

LP-MMSE Local Partial-Minimum Mean Square Error

LTE Long Term Evolution

MAC Medium Access Control

MAN Metropolitan Area Network

MCMC Markov Chain Monte Carlo

MIMO Multiple-Input Multiple-Output

mMIMO Massive Multiple-Input Multiple-Output

MMSE Minimum Mean Square Error

mmWave Millimeter-Wave

MR Maximum Ratio

MTBF Mean Time Between Failures

MTTA Mean Time to Absorption

MTTR Mean Time to Repair

NLoS Non-Line-of-Sight

NP Non-Protected

NPV Net Present Value

NR New Radio

OFDM Orthogonal Frequency-Division Multiplexing

OFDMA Orthogonal Frequency-Division Multiple Access

OLT Optical Line Terminal

O-RAN Open Radio Access Network

OPEX Operational Expenditures

P-RZF Partial-Regularized Zero Forcing

PD Partial Duplication

PDCP Packet Data Convergence Protocol

PDF Probability Density Function

PUE Power Usage Effectiveness

PHY Physical layer

P-MMSE Partial-Minimum Mean Square Error

PON Passive Optical Network

QAM Quadrature Amplitude Modulation

QoS Quality-of-Service

RAN Radio Access Network

RAM Random Access Memory

RB Resource Block

RF Radio Frequency

RLC Radio Link Control

RoC Radio-Over-Copper

RoF Radio-Over-Fiber

RoI Return on Investment

RRH Radio Remote Head

RRC Radio Remote Control

RU Radio Unit

Rx Radio Receiver

SB Segmented Bus

SC-FDMA Single-Carrier Frequency-Division Multiple Access

SoC System on Chip

SE Spectral Efficiency

SHR Self Healing Radio

SFP Small Form-factor Pluggable

SINR Signal-to-Interference-plus-Noise Ratio

SLA Service Level Agreement

SLA 40:90 Service Level Agreement where at least 40 % of the agreed UE rate is guaranteed to be achieved at anytime in 90 % of the coverage area

SNR Signal-to-Noise Ratio

SSD Solid-State Drive

SDMA Space Division Multiple Access

TCO Total Cost of Ownership

TDD Time Division Duplex

TRP Transmission-Reception Point

TRPU Transmission-Reception Point Unit

Tx Radio Transmitter

UatF Use-and-Then-Forget

UE User Equipment

UC User Centric

UMi Urban Micro

ULA Uniform Linear Array

UAV Unmanned Aerial Vehicle

VGA Variable Gain Amplifier

VR Virtual Reality

Wi-Fi Wireless Fidelity

List of Symbols

$\mathbf{A}$ Adjacency matrix of a graph

a_{deg} Maximum acceptable SE degradation due to quantized fronthaul samples in bps/Hz

$\mathbf{A}^{failed}$ Modified adjacency matrix of a graph representing connected equipment in a network due to a hardware failure

$A_{i,j}^{\text{Total}}$ Total accesses from state i to state j in cumulative analysis

$\mathcal{A}_{l,m}$ Set of UEs with TRPs in common with those served by TRP l at SB m

$a_{l,m,k}$ Weighting coefficients used in signal combining, aimed at optimizing spectral efficiency

B System radio bandwidth

B_{base} System radio bandwidth for network processing operations

B_c Coherence bandwidth

$b_{l,k}^{\text{data}}$ Bit width for the fronthaul data samples for TRPs in the SB l and UE k in the BTRP split

$b_{l,m'}^{\text{data}}$ Bit width for the fronthaul data samples in all antennas for TRP m at SB l in the BCPU split

$b_{l,m'}^{\text{pil}}$ Bit width for the fronthaul pilot samples for TRP m at SB l in the BCPU split

$\mathbf{B}_{l,m,k}$ Correlation error matrix of the channel estimates of the channel between TRP m at SB l and the UE k

$C_{a\&i}^{\text{CPU}}$ Acquisition and installation cost for CPU

$\mathcal{C}_i$ Cash flow at time i

$\mathcal{C}_j$ Set of users that TRPs j has channel estimates and/or coordinate interference

$\mathcal{C}_{l,m}$ Set of users that TRP m at SB l coordinate interference

C_i^{MkRw}	Markovian cost reward for state i
$C_{i,j}^{MkRw}$	Markovian cost reward for transition between states i and j
$C_{a\&i}^{\text{TRPs}}$	Acquisition and installation cost for TRPs
$C_{a\&i}^{\text{Xhaul}}$	Acquisition and installation cost for fronthaul interfaces
$C_{a\&i}^{\text{Xhaul}}$	Acquisition and installation cost for fronthaul interfaces
$C_{\text{fSpace}}^{\text{hourly}}$	Hourly costs of floor space
$C_{\text{rep}}^{\text{hourly}}$	Hourly costs with repairs
CAP_{bat}	CPU backup power battery capacity in Wh factoring in the depth of discharge
$\text{CAP}_{\text{Dcore}}$	GOPS capacity of the DSP processing core
CAP_{GPP}	GPP processing capacity in GOPS
CAP_{rack}	Maximum amount of GPPs that an edge cloud CPU rack can hold
$CC_{\text{all},t}^{\text{comb}}$	Required number of complex multiplications and divisions in the CPUs to generate the precoding vectors for all active UEs at time t
$CC_{l,m,t}^{\text{comb}}$	Required number of complex multiplications and divisions in TRP m at SB l to generate the precoding vectors for all active UEs at time t
$CC_{\text{all},t}^{\text{est}}$	Required number of complex multiplications and divisions in the CPUs to perform channel estimation for all active UEs at time t
$CC_{l,m,t}^{\text{est}}$	Required number of complex multiplications and divisions in TRP m at SB l to perform channel estimation for all active UEs at time t
$\mathbf{D}_k$	Diagonal binary matrix indicating which antennas concerning all TRPs serve UE k
$\mathcal{D}_j$	Set of users that TRPs j serves with data
$\mathcal{D}_l$	Set of users served by a SB
$\mathcal{D}_{l,m}$	Set of users that TRP m at SB l serves with data
$\mathbf{D}_{l,m,k}$	Diagonal binary matrix indicating which antennas in the TRP m at SB l serve UE k
$d_{l,m,k}$	Distance between TRP m at SB l and the UE k

$\mathbf{D}_{l,m,k}^{\text{modified}}$	Modified diagonal binary matrix indicating which antennas in the TRP m at SB l serve UE k due to failures
$\mathcal{D}_{l,\to m,t}$	Set of users by the SB starting from the m th TRP in the serial chain of connections at the UE load of time t
$\mathcal{D}_{l,\to m}$	Set of users by the SB starting from the m th TRP in the serial chain of connections
DS_k	Desired signal component for user k in the use-and-them-forget bound
$\mathbf{f}$	Binary array indicating which equipment are failed and functional
F_{cov}	Percentage of the SLA agreed rate equivalent to the network rate guarantee
F_{rate}	Percentage of the SLA agreed coverage area with at least the guaranteed rate
$F_{i,j}$	Frequency of accesses from state i to state j
$F_{l,m,t}$	Fronthaul bit rate to support an l, m TRP at a given time t
f_s	Sampling frequency
$\mathbf{F}_\theta$	Correlation matrix between the quantized signal and its original form
f_i	Frequency of accesses of state i
$\mathcal{F}_i$	Set of equiment failed in state i
$\mathbf{g}_{l,m,k}$	Multi-antenna NLoS channel between TRP m at SB l and the UE k in Rician model
$\text{GOPS}_{t,R}^{\text{CPU}}$	Number of GOPS to be executed at the CPU for an expected UE rate R at the UE load of the time t
$\text{GOPS}_{t,R}^{\text{CPUcommon}}$	GOPS requirements for CPU that are common in BCPU and BTRP splits
$\text{GOPS}_{t,R}^{\text{BTRP}_{l,m}}$	Number of GOPS of the specific operations in TRP m at acSB l of the BTRP split
$\text{GOPS}_{t,R}^{\text{TRP}_{l,m}}$	Number of GOPS to be executed in TRP m at acSB l for an expected UE rate R at the UE load of the time t
$\text{GOPS}_{t,R}^{BCPU}$	Number of GOPS of the specific CPU operations of the BCPU split

$\text{GOPS}_{\text{Chcd}}$ Reference GOPS value for channel coding network operations

$\text{GOPS}_{\text{HLct}}$ Reference GOPS value for higher-layer control operations

$\text{GOPS}_{\text{HLnt}}$ Reference GOPS value for higher-layer network operations

$\text{GOPS}_{\text{MpDp}}$ Reference GOPS value for layer mapping and demapping network operations

$\text{GOPS}_{\text{OFDM}}$ Reference GOPS value for OFDM modulation network operations

$\text{GOPS}_{t,R}^{\text{TRPcommon}}$ GOPS requirements for TRPs that are common in BCPU and BTRP splits

$\mathbf{h}_k$ Global channel from all antennas in all TRPs to UE k

$\mathbf{h}_{l,m,k}$ Multi-antenna channel between TRP m at SB l and the UE k

$\overline{\mathbf{h}}_{l,m,k}$ Multi-antenna LoS channel between TRP m at SB l and the UE k in Rician model

$\widehat{\mathbf{h}}_{l,m,k}$ Multi-antenna channel estimates between TRP m at SB l and the UE k

$\mathbf{H}_{\mathcal{S}_k}$ Aggregate channel matrix with channels of all UEs partially served by the same TRPs as UE k

$\mathcal{H}_{i \to j}$ Set of possible hardware that can fail when state j has one more failed equipment than state i

K Total number of users considered for the system model

$K_{l,m}$ Number of UEs connected to TRP m at SB l

IS_k Interferent signal component for user k in the use-and-them-forget bound

L Number of SBs

$\mathbf{L}(t)$ Array with the time spent at each after a time t in a CTMC

$\mathcal{M}$ Specific set of TRPs counts generating $\mathcal{R}$

$M_{act,t,C}$ Minimum number TRPs to support all UEs inside the coverage area at time t

$M_{act,t,R}$ Required number of TRPs to deliver a UE expected rate R at time t

$M_{act,t}$ Number of active TRPs to support the UEs load at time t

$\mathcal{M}_k$	Set of TRPs connected to UE k
$M_{tot,\max}$	Maximum number of possible deployed TRPs
M_{tot}	Total number of TRPs considered for the system model
$\text{MTBF}_{(\mathcal{A})}$	MTBF of the failure configuration equivalent to the set $\mathcal{A}$ of failed equipment
$\text{MTTR}_{(\mathcal{A})}$	MTTR of the failure configuration equivalent to the set $\mathcal{A}$ of failed equipment
$\text{MTTR}_{\min}$	Minimum MTTR of a set of failed equipment
M_l	Number of TRPs per SB l
$N_{\text{samples}}^{\text{daily}}$	Number of samples of time during the day
N_{DFT}	Dimension of the discrete Fourier transform
$N_{\text{GPPs},t}^{\text{act}}$	Number of active GPPs in the edge cloud CPU at time t
N_{GPPs}	Number of GPPs deployed at the edge cloud CPU
N_i	Number of equipment of type i
n_k^{dl}	Additive Gaussian noise at the signal received by UE k
$N_{l,m}$	Number of antennas in TRP m at SB l
$\mathbf{N}_{l,m}$	Receiver noise with i.i.d. elements
$\mathbf{n}_{l,m,t_k}$	Additive thermal noise for the pilot of the UE k received at the antennas of TRP m at SB l
$\breve{\mathbf{n}}_{l,m,t_k}$	Additive fronthaul quantization noise for the received pilot of the UE k in TRP m at SB l
$\mathbf{n}_{l,m}^{\text{ul}}$	Additive noise at the antennas of TRP m on SB l
N_{sc}	Number of subcarriers
$\mathbf{n}_{\text{ul}}$	Receiver noise across all antennas in all TRPs
$\mathbf{n}_{l,m}^{\text{ul}}$	Uplink additive receiver noise at the antennas of TRP m at SB l
N_{tech}^{i}	Number of technicians required for the repair of equipment of type i
N_{TRP}	Number if deployed TRPs

N_{ti}	Number of time intervals in the NPV analysis
N_{tot}	Total number of antennas considering all TRPs
$\mathbf{N}_{\mathbf{Y}_{l,m}}$	Additive quantization noise in the pilot final received at CPU trough fronthaul links
$\mathbf{o}$	Binary outage array that represents disconnection in the fronthaul links or TRPs due to component failure
$\mathcal{P}$	Set of equipment directly connected to the CPU
$P_{\text{CPU}}^{\text{peak}}$	Edge cloud CPU peak power consumption
$P_{\text{CPU},t}$	Edge cloud CPU power consumption at time t
$P_{\text{CPU},t}^{\text{cool}}$	Edge cloud CPU cooling system power consumption at time t
$P_{\text{CPU},t}^{\text{IT}}$	Edge cloud CPU IT equipment power consumption at time t
$P_{i \to j}$	Probability of going from a state i to a state j considering failure rate normalization
$\mathcal{P}_k$	Sets of UEs that use the same pilot as the user k
$p_{\text{LoS}(d_{l,m,k})}$	Probability of LoS communication between TRP m at SB l and the UE k based on distances
$p_{i,j}$	Transition probability from state i to state j at time t in a homogeneous DTMC
$p_{i,j}(t)$	Transition probability from state i to state j at time t in a CTMC
$p_{i,j,v}$	Transition probability from state i to state j at time interval v in a DTMC
$p_{(\mathcal{A})}^{0 \to t}$	Probability of failure after a time t concerning a system with equipment in the set $\mathcal{A}$
$\mathbf{P}$	Transition probability matrix for a homogeneous DTMC
$\mathbf{P}(t)$	Instantaneous state probability array for a CTMC at time t
$\mathbf{P}_{\mathcal{S}_k}$	Diagonal matrix with the transmit powers of all UEs partially served by the same TRPs as UE k
$\mathbf{P}_v$	Transition probability matrix for a DTMC at time interval v
P_{TRP}	TRP power consumption

P_n^{total}	Network power consumption at each time sample n
$P_{\text{Xhaul},t}$	Power associated with the backhaul/fronthaul network at time t
pr_{AFend}	TRP analog frontend price
pr_{ant}	Antenna price
pr_{bat}	Price for the battery's acquisition and installation in CPU deployment
pr_{DSP}	DSP price
$\text{pr}_{\text{DSP}}^{\text{base}}$	Fixed price related to other DSP construction parameter
pr_{Fdrop}	Price to install the final link from the FTTB infrastructure to the TRPs
$\text{pr}_{\text{FEport}}^{F_{l,\text{peak}}}$	Price of an Ethernet fronthaul switch port capable of supporting rates of $F_{l,\text{peak}}$
$\text{pr}_{\text{filter}}^{\text{ana}}$	Analog filter price
$\text{pr}_{\text{floor}}^{\text{year}}$	Price of renting per year per unit of area
pr_{GPP}	GPP acquisition price
pr_{IOint}	TRP I/O interface price
pr_{IQmod}	IQ modulator price
pr_{kWh}	Price of kilowatt-hour
$\text{pr}_{\text{rep}}^{i}$	Cost of replacement parts for a failure of equipment of type i
$\text{pr}_{\text{SFP}}^{F_{l,\text{peak}}}$	Price of a grey SFP capable of supporting rates of $F_{l,\text{peak}}$
$\text{pr}_{\text{Sinf}}^{\text{CPU}}$	Price of the support infrastructure for the edge cloud CPU
pr_{TRP}	TRP expected price
pr_{VGA}	VGA price
$\text{pr}_{\text{DAC\|ADC}}$	Price of DAC or ADC in the TRPs
$\text{pr}_{\text{rk\&nt}}$	Price for acquisition and installation cost of a rack and the network equipment in CPU deployment
PUE_{cool}	Edge cloud CPU cooling system power consumption at time t
pw_{ADC}	ADC power consumption
pw_{AFend}	TRP analog front-end power consumption

$\mathrm{pw}_{\mathrm{AFend}}$	TRP analog front-end power consumption
$\mathrm{pw}_{\mathrm{DAC}}$	DAC power consumption
$\mathrm{pw}_{\mathrm{DSP}}$	DSP power consumption
$\mathrm{pw}_{\mathrm{DSP}}^{\mathrm{other}}$	Non-GOPS dependant power consumption in the DSP
$\mathrm{pw}_{\mathrm{FEport}}^{F_{l,\mathrm{peak}}}$	Power consumption for an Ethernet fronthaul switch port capable of supporting rates of $F_{l,\mathrm{peak}}$
$\mathrm{pw}_{\mathrm{filter}}^{\mathrm{ana}}$	Analog filter power consumption
$\mathrm{pw}_{\mathrm{GPP}}^{\mathrm{idle}}$	Idle power consumption of the GPP
$\mathrm{pw}_{\mathrm{GPP}}^{\mathrm{peak}}$	Peak power consumption of the GPP
$\mathrm{pw}_{\mathrm{IOint}}$	TRP I/O interfaces power consumption
$\mathrm{pw}_{\mathrm{IQmod}}$	IQ modulator power consumption
$\mathrm{pw}_{\mathrm{Net}}^{\mathrm{rack}}$	Power consumption of the network equipment per rack in CPU deployment
$\mathrm{pw}_{\mathrm{SFP}}^{F_{l,\mathrm{peak}}}$	Power consumption of a grey SFP capable of supporting rates of $F_{l,\mathrm{peak}}$
$\mathrm{pw}_{\mathrm{Tx}}$	TRP transmission power
$\mathrm{pw}_{\mathrm{VGA}}$	VGA power consumption
q_k^{ul}	Uplink additive quantization noise affecting the uplink signal of user k
$\mathbf{q}_{l,m}^{\mathrm{dl}}$	Downlink transmission additive fronthaul quantization noise in TRP m at SB l
$q_{l,m,k}^{\mathrm{dl}}$	Downlink quantization noise of the user k in TRP m on SB l for the BTRP split.
$\mathbf{q}_{l,m}^{\mathrm{ul}}$	Uplink additive quantization noise associated with the quantized antenna signals in TRP m at SB l
$q_{l,m,k}^{\mathrm{ul}}$	Uplink additive quantization noise for the quantized combined signals of user k
$\mathbf{q}_{\theta}$	Quantization error concerning the original signal
$\mathcal{Q}_{\theta}(\mathbf{y})$	Quantization function with signal $\mathbf{y}$ as input
QN_k	Quantization noise component for user k in the use-and-them-forget bound

r	Discount rate for NPV
r_i	Reward of state i
$r_{i,j}$	Reward of the transition between states i and j
R	Expected UE rate
$\mathcal{R}$	Set of rates R for a specific set of TRP counts
R_{acov}	Average of the achievable UE rates higher or equal to $r_{F_{\text{cov}}}$
R_{agreed}	SLA agreed rate
R_{bcov}	Average of the achievable UE rates smaller than
$r_{F_{\text{cov}}}$	$(100 - F_{\text{cov}})$ th percentile rate in the UE achievable rate CDF
R_{s}	System reliability
$R_{(\mathcal{A})}^{0\to t}$	Reliability after a time t concerning a system with equipment in the set $\mathcal{A}$
R_k^{UatF}	User rate k under the use-and-them-forget bound
$\mathbf{R}_{l,m,k}$	Spatial correlation matrix that describes macroscopic propagation effects between TRP m at SB l and the UE k
$\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$	Covariance matrix of the quantized signal in its dimensions
$\mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}$	Covariance matrix capturing the relationship between the quantized signal and the original signal
$\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}}$	Covariance matrix capturing the relationship between the original signal and the quantized signal
$\mathbf{R}_{\mathbf{y}\mathbf{y}}$	Covariance matrix of the unquantized signal in its dimensions
$\mathcal{R}_{j,i}$	Set of equipment to be repaired from state j to i
S	Scenario area
s_{CPU}	Deployment area of the edge cloud CPU
$\mathcal{S}_k$	Set of UEs that are partially served by the same TRPs as UE k
s_{rack}	Necessary area to install a rack in CPU deployment
S_{tech}	

s_{TRP}	Area occupied by an TRP
$\mathrm{SE}_{\mathrm{base}}$	SE for reference GOPS values for network processing operations
$\mathrm{SE}_{t,R}$	SE for a expected UE rate R
SINR_k	SINR for user k
$\underline{\mathrm{SINR}}_k$	efecctive SINR for user k under the use-and-them-forget bound
$\mathbf{T}$	CTMC transition rate matrix
TCO_i	TCO at the period of time equivalent to i
T_n^{sample}	Duration of each time sample n in the day
$T_{\mathrm{GPP}}^{\mathrm{ins}}$	GPP installation time
T_i^{ope}	Time spend at the discrete time interval i
$T_{\mathrm{ope}}^{\mathrm{hours}}$	Total operational time in hours
T_i^{Total}	Total cumulative time at state i
$T_{\mathrm{ope}}^{\mathrm{sub}}$	Total operational time in the time unit of the subscription
T_{Pout}	Maximum duration of a power outage that can be managed
T_{rep}^i	Expected repair time of equipment of type i
$T_{\mathrm{TRP}}^{\mathrm{ins}}$	TRP installation time
T_{trv}	Technicians travel time
T_c	Coherence time
T_d	Delay spread time
T_i	Average permanence time at state i in a CTMC
t_k	Pilot index used for the user k
T_s	Symbol time
Tot_{hard}	Total amount of pieces of hardware in the system
U_{max}	Maximum number of TRPs that a UE can connect to
v	user velocity
$\mathbf{v}_k$	Linear combining array applied for user k concerning all antennas in all TRPs

$\mathbf{v}_k^{\mathrm{MMSE}}$	MMSE combining array for user k concerning all antennas in all TRPs
$\mathbf{v}_k^{\mathrm{P-MMSE}}$	Partial MMSE combining array for user k concerning all antennas in all TRPs
$\mathbf{v}_k^{\mathrm{P-RZF}}$	Partial RZF combining array for user k concerning all antennas in all TRPs
$\mathbf{v}_k^{\mathrm{MR}}$	MR combining array for user k concerning all antennas in all TRPs
$\mathbf{v}_{l,m,k}$	Linear combining array applied for user k in TRP m at SB l
$\mathbf{v}_{l,m,k}^{\mathrm{L-MMSE}}$	L-MMSE combining array for user k concerning all antennas in all TRPs
$\mathbf{v}_{l,m,k}^{\mathrm{LP-MMSE}}$	LP-MMSE combining array for user k concerning all antennas in all TRPs
$\mathbf{w}_k$	Linear precoding array applied for user k concerning all antennas in all TRPs
$\mathbf{w}_{l,m,k}$	Linear precoding array applied for user k from TRP m at SB l
$\mathbf{x}_{l,m}$	Transmitted signal of TRP m on SB l to the user k
$\hat{\mathbf{y}}$	Estimate of the quantized version of signal $\mathbf{y}$
$y_{\mathrm{CPU,k}}^{\mathrm{BTRP,ul}}$	Combined signal at the CPU corresponding to the symbol transmitted by the UE k
$\mathbf{y}_k^{\mathrm{ul}}$	received signal across all TRPs from user K
y_k^{dl}	Received data signal in a UE k
$\mathbf{y}_{l,m}^{\mathrm{ul}}$	The data signal received on the antennas of TRP m on SB l
$\mathbf{Y}_{l,m}^{\mathrm{pilot}}$	Received pilot signals in the TRP m at SB l
$y_{l,m}^{\mathrm{BTRP,ul}}$	Aggregated received signal available at the CPU from TRP m on SB l
$\tilde{\mathbf{Y}}_{l,m}^{\mathrm{pilot}}$	Fronthaul-quantized pilot signals in the TRP m at SB l
z_i	Mean time in spent at state i before absorption in a absorptive Markov chain
$\mathbf{Z}_{\mathcal{S}_k}$	Aggregated correlation error matrix of the channel estimates of all UEs partially served by the same TRPs as UE k considering a DCC framework transmssion

$\boldsymbol{\phi}_p$	pilot sequence p
$\boldsymbol{\Psi}_{l,m,t_k}$	Correlation matrix of the received pilot signal of user k in TRP m at SB l
$\mathcal{Z}_k$	Set of TRPs serving the UEs that are partially served by the same TRPs as UE k
α_{amp}	Expansion factor to account for losses in the amplification process
α_{SoC}	Price reduction factor due to SoC integration
α_{θ}	Empirical correlation coefficient between the quantized signal and its original form
$\tilde{\alpha}_{l,m}$	Linear correlation coefficient between unquantized and quantized pilot signal versions
$\alpha^{\text{dl}}_{l,m,k}$	Downlink linear correlation coefficient that quantifies the correlation between the original and quantized uplink signals for user k from TRP m at SB l
$\alpha^{\text{ul}}_{l,m,k}$	Uplink linear correlation coefficient that quantifies the correlation between the original and quantized uplink signals from user k in TRP m at SB l
α_t	Active UE load ratio at time t
$\bar{\theta}$	Elevation angle of a multipath channel component
$\bar{\varphi}$	Azimuth angle of a multipath channel component
$\beta_{l,m,k}$	Large scale channel gains between TRP m at SB l and the UE k
η_k	Uplink transmit power of UE k
γ_{Chcd}	Scaling factor for channel coding network operations
γ_{HLct}	Scaling factor for higher-layer control operations
γ_{HLnt}	Scaling factor for higher-layer network operations
γ_{inv}	Price slope for inverter in the Edge CPU
γ_{MpDp}	Scaling factor for layer mapping and demapping network operations
γ_{OFDM}	Scaling factor for OFDM modulation network operations
γ_{prDcore}	Price slope for the necessary number of cores in the DSP

γ_{pwDcore}	Power slope related to the DSP idle core operation
γ_{pwDSP}	Power slope related to the operations in all cores of the DSP
$\gamma_{\text{Co}\|\text{PD}}$	Price slope for CPU cooling and power distribution infrastructure
$\kappa_{l,m,k}$	Rician factor between TRP m at SB l and the UE k
λ	System failure rate
λ_{sub}	subscription price of user
$\lambda_{i,j}$	Transition rate from state i to state j in a CTMC
ϕ	Prelog factor for spectral efficiency calculations
$\boldsymbol{\pi}$	Array with steady-state probabilities in a Markov chain
π_i	Steady-state probabilities of state i in a Markov chain
ρ_K	UE density
σ_φ	Angular standard deviation of a multipath channel components
$\sigma^2_{\text{ul}\|\text{dl}}$	Addtive thermal noise component for user k in the use-and-them-forget bound
τ_d	Number of samples inside the coherence block used for downlink
τ_{dp}	downlink orthogonal pilot sequence length
τ_p	Number of pilot samples per coherence block
τ_u	Number of samples inside the coherence block used for uplink
τ_{up}	uplink orthogonal pilot sequence length
τ_c	Coherence block number of samples
ϱ_k	Downlink power transmission coefficient to the user k
$\varrho_{l,m,k}$	Downlink power transmission coefficient to the user k in TRP m at SB l
ς_k^{dl}	Downlink signal intended for UE k
ς_k^{ul}	Uplink transmitted signal from user k
Ω	Space state set in a Markov chain
Ω_A	Absorptive state set in a absorptive Markov chain
Ω_T	Transient state set in a absorptive Markov chain

List of publications

Publications directly related to the thesis content and proposals:

A. Papers:

A-1. L. S. Furtado, A. L. P. Fernandes, A. A. Ohashi, F. S. Farias, A. M. Cavalcante and J. C. W. A. Costa, "Cell-Free Massive MIMO Deployments: Fronthaul Topology Options and Techno-Economic Aspects," 2022 16th European Conference on Antennas and Propagation (EuCAP), Madrid, Spain, 2022, pp. 1-5, doi: 10.23919/EuCAP53622.2022.9768969.

A-2. A. L. P. Fernandes, D. D. Souza, D. B. da Costa, A. M. Cavalcante and J. C. W. A. Costa, "Cell-Free Massive MIMO With Segmented Fronthaul: Reliability and Protection Aspects," in IEEE Wireless Communications Letters, vol. 11, no. 8, pp. 1580-1584, Aug. 2022, doi: 10.1109/LWC.2022.3166485.

A-3. A. L. P. Fernandes, D. D. Souza, C. Natalino, F. Tonini, A. M. Cavalcante, P.Monti and J. C. W. A. Costa, "A Cost Assessment Methodology for User-Centric Distributed Massive MIMO Architectures," in IEEE Open Journal of the Communications Society, vol. 5, pp. 3517-3543, 2024, doi: 10.1109/OJCOMS.2024.3406374.

B. Patents:

B-1. A. L. P. Fernandes, L. S. Furtado, R. M. Rodrigues, J. C. W. A. Costa, G. S. Borges, A. M. Cavalcante, M. V. Marquezini and I. Almeida, “Self-healing method for fronthaul communication failures in serial cell-free networks“, PCT/IB/059016, Filing date: Sep. 2021.

Publications related to cell-free mMIMO (indirectly related to the thesis proposals):

C. Papers:

C-1. A. A. Ohashi, D. B. da Costa, A. L. P. Fernandes, W. Monteiro, R. Failache, A. M. Cavalcante et al., “Cell-Free Massive MIMO-NOMA Systems With Imperfect SIC and Non-Reciprocal Channels,” IEEE

Wireless Communications Letters, vol. 10, no. 6. Institute of Electrical and Electronics Engineers (IEEE), pp. 1329–1333, Jun. 2021. doi: 10.1109/lwc.2021.3066042.

C-2. R. V. P. Failache, A. A. Ohashi, A. Fernandes, R. M. Rodrigues, A. M. Cavalcante, and J. W. Costa, “A Fair Comparison Between OMA and NOMA For Cell-Free Massive MIMO Systems,” Anais do XXXIX Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Sociedade Brasileira de Telecomunicações, 2021. doi: 10.14209/sbrt.2021.1570724206.

C-3. M. M. M. Freitas, D. D. Souza, A. L. P. Fernandes, D. B. Da Costa, A. M. Cavalcante, L. Valcarenghi et al., "Scalable User-Centric Distributed Massive MIMO Systems with Limited Processing Capacity," ICC 2023 - IEEE International Conference on Communications, Rome, Italy, 2023, pp. 4298-4304, doi: 10.1109/ICC45041.2023.10279694.

C-4. T. T. R. Ueoka, M. A. S. Costa, D. D. Souza., M. M. M. Freitas, A. L. P. Fernandes, G. S. Borges et al„ “Pilot Allocation and Assignment Optimization in User-Centric Distributed Massive MIMO Networks,” Anais do XLI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Sociedade Brasileira de Telecomunicações, 2023. doi: 10.14209/sbrt.2023.1570923555

C-5. M. A. S. Costa, M. M. M. Freitas, T. T. R. Ueoka, D. D. Souza., A. L. P. Fernandes, A. M. cavalcante et al, “A Fairer Comparison of Processing Implementations in User-Centric Distributed Massive MIMO Systems,” Anais do XLI Simpósio Brasileiro de Telecomunicações e Processamento de Sinais. Sociedade Brasileira de Telecomunicações, 2023. doi: 10.14209/sbrt.2023.1570915442.

Additional publications during PhD. studies (unrelated to the thesis):

D. Papers:

D-1. M. V. P. Oliveira, A. L. P. Fernandes, F. V. S. dos Reis, H. P. Kuribayashi, M. A. de Souza, L. N. Goncalves et al., "Technical-Economic Analysis of Photovoltaic Systems Applied to Heterogeneous Mobile Networks in Rural Areas," in IEEE Latin America Transactions, vol. 18, no. 11, pp. 1900-1908, November 2020, doi: 10.1109/TLA.2020.9398631.

D-2. D. Acatauassu, M. Licá, A. A. Ohashi, A. L. P. Fernandes, M. M. M. Freitas, J. C. W. A. Costa et al., "An Efficient Fronthaul Scheme Based

on Coaxial Cables for 5G Centralized Radio Access Networks," in IEEE Transactions on Communications, vol. 69, no. 2, pp. 1343-1357, Feb. 2021, doi: 10.1109/TCOMM.2020.3039860.

Notations

Boldface lowercase and uppercase letters denote vectors and matrices, respectively, and $diag(\,.\,)$ is operator that extracts the diagonal elements of a square matrix to form a diagonal matrix. The superscripts $(\cdot)^{\mathrm{H}}$ and $(\cdot)^{\mathrm{T}}$ denotes the conjugate-transpose and transpose operations. The identity matrix is $\mathbf{I}_N$ has dimensions $N \times N$. The cardinality of the set $\mathcal{A}$ is represented by $|\mathcal{A}|$. The trace, euclidean norm, and expectation operator are denoted as $\mathrm{tr}(\,.\,)$, $\|\,.\,\|$, and $\mathbb{E}\{\,.\,\}$, respectively. The notation $\mathcal{CN}(\mu, \sigma^2)$ stands for a complex Gaussian random variable with mean μ and variance σ^2. The notation $\mathcal{CN}(\mu, \mathbf{A})$ represents a complex random vector with mean μ and covariance matrix given by a $\mathbf{A}$.

1 Introduction

Mobile communication networks have experienced continuous growth in user demands over the years, driving consistent expansions in their capacities. From supporting just analog voice in circuit-based First Generation of Mobile Systems (1G), these networks have evolved into digital packet-based systems, capable of delivering user data rates exceeding 10 Gbps in the Fifth Generation of Mobile Systems (5G). Over these five generations, the capacity evolution was supported by making the Transmission-Reception Point (TRP) distribution denser, increasing the available bandwidth, and utilizing the spectrum more efficiently. Despite this, there is no defined limit for the amount of capacity that will ever be needed as each increase in capacity generates new application possibilities, which in turn stimulate a subsequent rise in demand that requires additional capacity. In this context, improvements in network density, available bandwidth, and Spectral Efficiency (SE) are essential to support next-generation systems(Björnson; Hoydis; Sanguinetti, 2017)(Marzetta, 2015)(Zhang et al., 2020).

All mobile communication generations until nowadays have been based on the cellular paradigm, where the network coverage area is divided into cells, each served by a TRP, also called Base Station (BS) for the cellular context. In this structure, users within a given cell communicate exclusively with their respective BSs, and each cell operates as a isolated network. Along the continuous evolution of mobile systems, the cell density and available bandwidth are approaching a saturation point where further improvements are complicated and expensive. In this way, increasing SE is of fundamental importance to future networks. Massive Multiple-Input Multiple-Output (mMIMO) systems have BSs with a large number of antennas, typically 64 or more, serving single-antenna User Equipments (UEs). These systems can increase SE using computationally cost-affordable linear combiners/precoders without increasing transmission power, using complex transceivers, or using advanced processing techniques in UEs side (Björnson; Hoydis; Sanguinetti, 2017)(Marzetta, 2015)(Zhang et al., 2020).

Despite the benefits of mMIMO, there are performance limitations associated not with the technology itself but with the cellular operational paradigm. At the cell borders, user interference tends to be higher, resulting in non-uniform data rates across the coverage area. The problem is even worse in denser distributions of TRPs, which is a key strategy for increasing data capacity. As TRPs are positioned closer together, there is less decay in interfering signals, potentially increasing inter-cell interference and reducing user data rates. Thus, while mMIMO can substantially increase SE for users near a cell's BS, it tends to perform poorly in terms of SE for users near cell borders due to this inherent inter-cell interference. In this manner, the SE gains from mMIMO will always be suboptimal as long the cellular paradigm remains used as the basis of the system (Björnson; Sanguinetti, 2020b)(Demir; Björnson; Sanguinetti, 2021).

A more optimal implementation approach is to co-process and transmit user signals using multiple TRPs, providing multiple data channels to each UE that are spaced by much more than the wavelength of the transmission. This property, known as macro-diversity, enhances the system's resilience to signal fading and interference. The co-processing and co-transmission approach creates a distributed mMIMO system that eliminates the TRPs densification issue for future mobile networks while taking advantage of the increased SE of mMIMO systems. Besides that, the TRP does not need a considerable number of antennas in this distributed scheme. In essence, each user is connected to a large number of antennas, as they are connected to multiple TRPs (Björnson; Sanguinetti, 2020b)(Demir; Björnson; Sanguinetti, 2021). For example, a 64-antenna connection can be achieved by a user connected to eight TRPs with eight antennas.

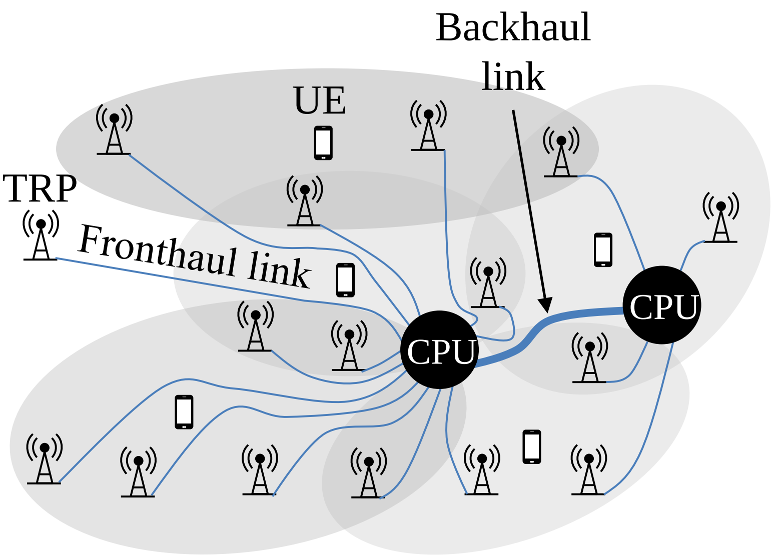
1.1 Cell-free mMIMO

A cell-free mMIMO network is a way to implement a distributed mMIMO system that effectively eliminates mobile network cell boundaries. It employs a large number of TRPs scattered across a coverage area, each equipped with one or more antennas. The TRPs jointly processes user signals by interchanging information via fronthaul links between themselves and one or more edge-cloud Central Processing Units (CPUs). These CPUs are connected between themselves and the operator backbone through backhaul links and facilitate the information exchange and orchestrate the system's overall coordination (Demir; Björnson; Sanguinetti, 2021)(Interdonato et al., 2019)(Ngo et al., 2017).

Figure 1 illustrates the architecture of a cell-free mMIMO network, depicting the components mentioned before. The TRPs serving each UE are depicted as gray ellipses. For scalability reasons, not all TRPs serve all UEs, but from the user's perspective, connectivity is still provided by a large number of distributed antennas. Under this architecture, a higher and fairer SE can be achieved even in a dense TRP deployment, transforming cell-free mMIMO network in an essential technology for next-generation mobile communication systems (Zhang et al., 2020). Their adoption can be beneficial in scenarios requiring high data rates with consistent service coverage, including outdoor urban macro areas, hotspot areas, as well as indoor offices and factories (Interdonato et al., 2019).

The telecommunications industry has recently presented the first practical solution for cell-free mMIMO networks (Interdonato et al., 2019). An example is Ericsson Radio Stripe System (ERSS), which is based on a fronthaul where TRPs are sequentially

Figure 1 – Illustration of the architecture of cell-free mMIMO network. Fronthaul links connect an edge cloud CPUs to the TRPs. Backhaul links interconnect CPUs. UEs are served by a limited optimal set of TRPs with available resources.



Source: elaborated by the author.

connected in series (Interdonato et al., 2019). This configuration, also called segmented fronthaul cell-free mMIMO, can increase the deployment scalability by large margins, but it has a potential reliability issue. A failure in one fronthaul segment or TRP can lead to sequential transport communication failures, resulting in data outages across all subsequent TRPs and fronthaul segments.

A typical approach to compensate the hardware failure effects in communication networks is the utilization of protection/redundancy schemes. Specifically, in the context of mobile networks, some works have already employed solutions for failure recovery, such as in Selim et al. (2016) and Selim and Kamal (2018), which make use of technologies like Self Healing Radio (SHR) and Unmanned Aerial Vehicle (UAV) to provide an alternative wireless fronthaul and secondary mobile BSs, respectively. Despite the existence of these protection and redundancy schemes, there is a scarcity of analyses from the cell-free mMIMO communication paradigm perspective. Therefore, conducting technical analysis to evaluate the impact of fronthaul communication failures on cell-free mMIMO, particularly when operating under segmented fronthaul with or without protection strategies, raises a compelling research topic.

Another important aspect of the ERSS industry solution is that it is an integrated system, so the TRPs and fronthaul segments are combined into a single device using a specific unified communication structure that should be considered when developing any protection scheme. In this way, protection schemes for non-integrated and integrated cell-free mMIMO systems will differ due to the specific characteristics of the latter.

In addition to the initial industry solutions, academia is also evolving the cell-free mMIMO concept by providing advancements in modeling for channel behavior, computational requirements, fronthaul signaling, migration from the cellular paradigm, and integration with existing standards (Björnson; Sanguinetti, 2020b)(Demir et al., 2024)(Femenias; Riera-Palou, 2020)(Interdonato et al., 2019)(Kim et al., 2022)(Polegre et al., 2020)(Demir et al., 2024).

These advancements are significant in clarifying how the system will be deployed, as cell-free mMIMO was previously a theoretical concept with unclear modeling of the system's diverse requirements and capacities, i.e., it was unclear how to deploy the system effectively. Despite these models, there is still no clear answer on the most feasible way to deploy the system, since comprehensive techno-economic analysis and the economic feasibility for cell-free mMIMO networks have not been thoroughly analyzed in the literature. Although some initial studies have been conducted, they neglect essential operational aspects like user demands, fronthaul limitations, and realistic hardware processing capacities and requirements (Xiao; Mähönen; Simić, 2022)(Xiao; Mähönen; Simić, 2023).

A comprehensive techno-economic model can help answer the discussion between centralized and distributed processing implementations in the literature. The architecture of cell-free mMIMO is inherently distributed. However, the processing implementation can be either centralized or distributed. This flexibility arises from performing baseband functions locally at the TRPs or the edge CPU. When tasks such as channel estimation and precoding computation occur at the TRPs, the processing is distributed, utilizing simpler combining/precoding techniques like maximum ratio. This approach aligns with the distributed nature of cell-free mMIMO and offers high computational resource efficiency. Conversely, centralized processing, which involves performing these tasks at edge CPUs, enables more advanced processing techniques, potentially achieving superior performance at the expense of increased computational complexity (Björnson; Sanguinetti, 2020a)(Femenias; Riera-Palou, 2020).

Initially, cell-free mMIMO was mainly based on distributed processing due to its simplicity, which was believed to increase the system's scalability and reduce fronthaul signaling (Ngo et al., 2017)(Interdonato et al., 2019). However, it was eventually proved that centralized processing could also be scalable and potentially have lower fronthaul signaling than the distributed case while providing much higher performance (Björnson; Sanguinetti, 2020a). Nevertheless, this does not mean centralized approaches are always superior to distributed ones. The computational complexity can be orders of magnitude higher than the distributed case (Freitas et al., 2023). Besides that, the fronthaul requirements can be similar to the distributed case depending on the antenna count on the TRP, its supported number of users, and the supported user data rate. Then a comprehensive techno-economic comparison is essential to appropriately assess the superiority of centralized or distributed processing in different situations.

In this context, practical aspects that are critical for a feasibility analysis are still ignored by proposed solutions for cell-free mMIMO networks, including cost affordability and hardware reliability analysis. This thesis explores these avenues, aiming to devise solutions applicable to both urban and indoor environments.

1.2 Related works

This section presents a set of works developed by the literature involving cell-free mMIMO related to the technology development in a practical sense, its cost affordability, and hardware reliability analysis.

1.2.1 Cell-free mMIMO

The notion of co-processing signals from users across multiple TRPs emerged in parallel with mMIMO to support the increasing demands of Fourth Generation of Mobile Systems (4G) and 5G. This idea has been propagated under various designations, including network Multiple-Input Multiple-Output (MIMO), Coordinated MultiPoint-Joint Transmission (CoMP-JT), distributed MIMO, and multi-cell MIMO cooperative networks. Fundamentally, the co-processing in these systems can be implemented in two primary ways: network-centric or User Centric (UC). The network-centric approach groups TRPs into separate clusters, with each cluster jointly transmitting to the UEs in their shared coverage area. While this simplifies cooperative deployment, it can limit performance gains since users are not always served by their optimal TRPs set, leading to potential service quality issues at cluster borders (Björnson; Sanguinetti, 2020b)(Interdonato et al., 2019).

On the other hand, the UC approach ensures that users receive transmissions from an optimized subset of TRPs that maximize performance. This subset considers TRP connection limits, channel conditions, and fronthaul capacity for signal sharing. Even non-serving TRPs might adjust for potential interference to an user. This strategy virtually eliminates cell boundaries, offering superior interference management compared to the network-centric method (Björnson; Sanguinetti, 2020b)(Interdonato et al., 2019).

As mMIMO emerged as a primary solution for enhancing SE in 5G, coordinated transmission strategies were somewhat overshadowed by its prominence. However, as mMIMO matured, the emphasis shifted to coordinated transmission techniques, mainly due to their superior interference cancellation capabilities. This resurgence came under the banner of cell-free mMIMO, effectively a user-centric distributed MIMO system employing many TRPs but built upon the foundation of technologies crafted for cellular mMIMO (Björnson; Sanguinetti, 2020b)(Ngo et al., 2017).

The initial works Ngo et al. (2015), Ngo et al. (2017), and Ngo et al. (2018b) focused on deriving closed-form expressions for the achievable rate using a conjugate beamforming precoder, which could be easily implemented in a distributed way and using power control strategies that maximized the minimum rate in the network or energy efficiency of the system, which were popular power control strategies for cellular mMIMO. The rate or energy efficiency was then compared to one achieved on cellular small-cell networks, showing that a cell-free mMIMO system could significantly outperform small-cell cellular networks, being capable of providing throughput and network densification needed for future networks. Despite this, various practical aspects were disregarded in these works. The main one was the assumption that all TRPs connect to all UEs, which would be impossible in a wide-area networks. As the cell-free mMIMO concept was initially defined under this assumption, the literature sometimes calls it as canonical cell-free mMIMO.

In Interdonato et al. (2019), a comprehensive examination of the necessary practical aspects for a realistic assessment of a cell-free mMIMO system was undertaken, commenting on factors such as transport infrastructure, channel estimation, pilot assignment, signal processing, and power control. A critical concern revolved around system scalability, which is fundamental to ensure the deployment of dense networks like cell-free mMIMO . Addressing this, Interdonato, Frenger and Larsson (2019) unveiled a fully distributed, UC, and scalable architecture ensuring scalability across signal processing, network topology, and power control. While centralized processing strategies might have offered superior performance, Interdonato, Frenger and Larsson (2019) and Interdonato et al. (2019) favored intelligent distributed processing methodologies to avoid fronthaul signaling overload and to safeguard the system's scalability. This advocacy for distributed strategies stemmed from the prevailing belief that the naturally distributed architecture of cell-free mMIMO could provide excellent performance with simple conjugate beamforming precoders, which are inherently scalable. Then, more advanced or centralized signal processing was deemed unnecessary, even with the gain they would provide.

In Björnson and Sanguinetti (2020a), the performance of cell-free mMIMO under centralized and distributed signal processing implementation was thoroughly analyzed. Centralized processing delegated channel estimation and precoding computation to CPUs, while distributed processing assigned these tasks to TRPs. The works also compared cell-free mMIMO against a small-cell operation and explored precoders beyond conjugated beamforming. Findings indicated that cell-free mMIMO did not consistently outperform small-cell systems when relying solely on distributed conjugate beamforming. Despite this, a more advanced Local-Minimum Mean Square Error (L-MMSE) precoder could always surpass small cells for distributed processing. Besides that, centralized processing utilizing an Minimum Mean Square Error (MMSE) precoder was pointed out as the optimal way to operate cell-free mMIMO, because it offered superior SE and could potentially demand less fronthaul signaling load than its distributed counterparts. This fact arose because each TRP was expected to serve more users than its number of antennas, and the fronthaul load scaled with the user and TRP antenna counts for distributed and centralized processing, respectively. Nevertheless, the study recognized that a non-infinite precision fronthaul, with adequate representation of the bit-width used for the complex scalars sent through the fronthaul, could potentially change this conclusion.

Another essential practical implementation aspect investigated by Interdonato et al. (2019) was the fronthaul cabling complexity and possible deployment scalability issues it could bring. The work proposed a segmented fronthaul structure, where TRPs were connected serially to each other, to tackle this problem. According to Björnson and Sanguinetti (2020a), this serial fronthaul configuration was advantageous for distributed processing. It typically featured fewer users per serial chain compared to the number of antennas in each chain. Consequently, distributed processing led to reduced fronthaul signaling, even when using the same bit-width for the complex scalars transmitted across the fronthaul in both centralized and distributed processing schemes. Besides that, in the solution proposed by Interdonato et al. (2019), the TRPs were integrated into a equipment known as radio stripe, where the TRP data transfer interconnection was made via a shared bus that also provided synchronization and power supply. This integrated solution, known as ERSS, aimed to reduce even more problems with the deployment complexity, limited capacity of back/fronthaul connections, and network synchronization.

The utilization of serially interconnected TRPs in cell-free mMIMO was further investigated in Miretti, Björnson and Gesbert (2021) and Shaik, Björnson and Larsson (2020), where sequential unidirectional processing for uplink and downlink was evaluated. In these works, each TRP sends its local estimated signals, Channel State Information at the Transmitter (CSIT), and error statistics to its neighbor TRP closer to the CPU in terms of fronthaul hops. Then, each TRP used its own information and the one from other TRPs to process and decode their signals. This structure formed a compute-and-forward architecture, resulting in a cell-free mMIMO network with local processing and more information in each TRP than just their channel knowledge, i.e., it provided a hybrid between a centralized and distributed processing that could only work in serially connected systems, increasing the throughput of cell-free mMIMO networks with segmented fronthaul operating under decentralized processing. The results showed that sequential local processing performed much better than a purely distributed processing approach, with a throughput that was only 12% smaller than the high-performance centralized processing approach.

All previously mentioned works, while advancing practical deployment considerations for cell-free mMIMO, were still considering an infinite precision fronthaul. In this context, Bashar et al. (2018) and Bashar et al. (2019) addressed non-infinite precision fronthaul, focusing on modeling quantized signal impacts on Signal-to-Interference-plus-Noise Ratio (SINR) and only on the uplink operation. However, these studies did not considered a centralized processing implementation analogous to the optimal one in Björnson and Sanguinetti (2020a), which has a different representation for the number of complex scalars sent through the fronthaul. Subsequently, Femenias and Riera-Palou (2020) expanded upon this by examining the implications of quantized signals on cell-free mMIMO's performance across uplink and downlink operations. This study modeled quantization-related errors using an Additive Quantization Noise Model (AQNM) based on Bussgang decomposition and presented models for two functional splits representing distributed and centralized processing implementations. These models aligned with those in Björnson and Sanguinetti (2020a), accommodating a variable bit-width depending on user numbers and fronthaul capacity. The results corroborated with Björnson and Sanguinetti (2020a), proving that fronthaul signaling would be potentially way smaller in the centralized implementation for a similar level of user rate performance.

In Ngo et al. (2018a), the performance of cell-free mMIMO in a Ricean fading channel, comprising Line-of-Sight (LoS) and Non-Line-of-Sight (NLoS) components, was examined. The study highlighted that dense deployment leads to a dominant LoS component, contrasting with previous literature that largely relied on the Rayleigh fading model. This latter model fits best to the scenarios with strong scattering and weak LoS components, affecting the design of channel estimators and the subsequent performance of processing methods. Polegre et al. (2020) offers a more comprehensive channel modeling by factoring in correlated NLoS components. The findings indicated that the predominant LoS aspect of Ricean fading enhanced channel hardening, which was known to be much weaker in cell-free mMIMO compared to cellular mMIMO, resulting in superior rates than those results obtained by the Rayleigh model.

In Björnson and Sanguinetti (2020b), the cell-free mMIMO's scalability was re-examined, encompassing more than just the distributed processing considered in Interdonato, Frenger and Larsson (2019). The work defined scalability as maintaining finite processing requirements, fronthaul/backhaul signaling, and total power, even as user count grew indefinitely. To achieve this behavior, scalable strategies to form user-centric TRPs clusters and pilot assignments were introduced, which used a Dynamic Cooperation Clustering (DCC) framework combined with an initial access procedure. This ensured scalable channel estimation. Moreover, Partial-Minimum Mean Square Error (P-MMSE) and Local Partial-Minimum Mean Square Error (LP-MMSE) combiners were proposed as scalable solutions for centralized and distributed processing implementations. The results proved that centralized processing could also be implemented in a scalable way. Expanding this framework, Demir, Björnson and Sanguinetti (2021) introduced scalable power allocation methods and a centralized Partial-Regularized Zero Forcing (P-RZF) combiner. Despite this, Freitas et al. (2023) challenged the scalability definition from Björnson and Sanguinetti (2020b), highlighting that indefinitely increasing TRPs might reintroduce non-scalability. Accordingly, Freitas et al. (2023) modified the approach to guarantee scalability under such conditions.

Recent developments in cell-free mMIMO research have investigated its practical aspects, including considerations on transitioning from traditional cellular mMIMO infrastructures and the integration into practical specifications and guidelines such as the Open Radio Access Network (O-RAN) architecture Kim et al. (2022)(Demir et al., 2024). Besides that, the developments also included modeling for system coordination, fronthaul signaling, and computational complexities of processing procedures (Demir; Björnson; Sanguinetti, 2021)(Demir et al., 2024). This trend represents a shift in literature, and cell-free mMIMO is starting to be seen as something beyond a theoretical concept. As such, the development of feasibility evaluation frameworks, to evaluate literature-developed solutions in terms of reliability and economic impact, can be a crucial contribution to the field.

1.2.2 Reliability of cell-free mMIMO with segmented fronthaul

In Interdonato et al. (2019), it is suggested that the inherent design of cell-free networks, which feature a much higher number of TRPs compared to the number of UEs, enhances system reliability. This happens because even if several TRPs fail, many still remain functional and can serve as alternatives. However, this conjecture requires detailed analysis to be substantiated, especially for segmented fronthaul deployments. In such setups, failures of fronthaul segments or TRPs can cause data outages across a large number of TRPs, potentially reducing system performance and reliability in a significant way.

Prior to the research derived from this thesis, the literature lacked a robust reliability analysis for cell-free mMIMO systems with segmented fronthaul. However, other works have developed solutions for failure recovery in fixed access and cellular mobile networks. For example, Selim et al. (2016) and Selim and Kamal (2018) discussed using SHR and UAV for alternative wireless fronthaul and secondary mobile BSs, respectively. These approaches demonstrated that wireless fronthaul rerouting and TRP failure compensation were vital for enhancing reliability in cellular networks. Moreover, Fernandez and Stol (2016) proposed a risk assessment approach to evaluate the costs and client dissatisfaction in Passive Optical Networks (PONs) deployments as fixed Fiber to the Home (FTTH) access networks under various protection schemes, including equipment duplication and cross-connection.

Although the protection strategies in Selim et al. (2016) and Fernandez and Stol (2016) demonstrated advantages in fixed and cellular access networks, these strategies were designed outside of the scope of a cell-free mMIMO communication. Nevertheless, these schemes can serve as an inspiration or as a basis for newer ones. For instance, the models in Fernandez and Stol (2016) can serve as a basis for identifying crucial equipment that should be cross-connected or duplicated.

Additionally, the Markov Chain Monte Carlo (MCMC) approach used in Fernandez and Stol (2016) and Fernandes (2019) modeled software and hardware failures over time, capturing the interdependence between different equipment and the potential for multiple failures. This model could be a powerful tool for analyzing cumulative failures in cell-free mMIMO systems, which could be relevant as individual equipment failures might not have always required repairs given the various TRP alternates unless a major failure occurs. In fact, it may be cost-effective to allow equipment to fail up to a certain threshold of performance degradation before conducting repairs, optimizing resource allocation, and maintenance scheduling.

1.2.3 Techno-economics of cell-free mMIMO networks

In Oughton and Lehr (2022), an extensive analysis of literature concerning 5G techno-economics was carried out. The primary aim of the investigation was to provide recommendations for techno-economic assessments of future next-generation mobile communication systems, like cell-free mMIMO systems. The study reached several essential conclusions. Firstly, the accuracy and reliability of any techno-economic analysis hinged on a well-defined network dimensioning procedure. Secondly, when evaluating financial metrics, it was deemed imperative to consider both Capital Expenditures (CAPEX) and Operational Expenditures (OPEX) over a defined deployment period. It was observed that this approach offered a clearer understanding of the cost-benefit ratio of different technologies. Lastly, a sensitivity analysis was considered essential to ascertain the validity of any proposed models and methods. Such analysis was seen as the ideal way to explain the dependence of modeling results on essential input parameters and modeling decisions.

Despite mMIMO being pointed out as an integral enabler of the 5G Zhang et al. (2020), its techno-economics still needed to be explored by the literature, especially for its distributed user-centric implementation. Most of the works in the expansive techno-economic review of Zhang et al. (2020) did not focus on mMIMO, just pointing it out as a way to increase the throughput of the cell. The ones that did focus on mMIMO considered just specific use cases. An example is Jha and Saha (2018), which assessed the profitability of mMIMO operating under Millimeter-Wave (mmWave) for the state of Texas in the United States. The analysis was valid but disregarded one of the main advantages of mMIMO, the ability to massively increase SE in sub-6 GHz bands, enhancing the user experience for the coverage layer of the cellular network (Björnson; Hoydis; Sanguinetti, 2017).

Simplified techno-economic models on cell-free mMIMO systems were introduced by recent literature not covered by Oughton and Lehr (2022). In Xiao, Mähönen and Simić (2022), a comparative analysis of cost efficiency was presented, evaluating cell-free mMIMO against small cells. The study investigated various sizes of TRP clusters for each UE and examined different fiber transport connections under single and multiple CPU scenarios. The findings suggested that cell-free mMIMO could achieve superior throughput at a reasonable system cost, contingent on carefully chosen cluster sizes and inter-CPU cooperation levels. The study model was adequate for the proposed analysis but had several shortcomings for further development. These include the absence of OPEX modeling, reliance on only a centralized non-scalable MMSE precoder, use of a fixed TRP quantity, and simplified step models for the costs associated with deploying TRPs and CPUs. For the latter case, the calculations scaled solely with the size of the subset of TRPs serving each UE.

One of the primary limitations of Xiao, Mähönen and Simić (2022) was addressed by Xiao, Mähönen and Simić (2023), which expanded the analysis to incorporate energy-related OPEX into the model. Nevertheless, this subsequent work did not address the other deficiencies in the initial model. Furthermore, essential types of OPEX were still not explored. Although the energy model could be considered adequate, there was room for expansion, as many computational operations at the CPU and TRP were overlooked.

Additionally, neither Xiao, Mähönen and Simić (2022) nor Xiao, Mähönen and Simić (2023) addressed the dimensioning of the necessary number of TRPs concerning demands. Instead, these works circumvented this challenge by assuming a fixed number of TRPs and delved into other aspects, like transport network configuration. Moreover, it is evident that existing literature's dimensioning procedures for mobile systems, like the ones in Yaghoubi et al. (2018), are focused on the cellular paradigm and not adept at determining the required number of TRPs, because of the multiple coordinated TRP connections to a single UE. In this context, dimensioning procedures in economical analysis for cell-free mMIMO is imperative for future works, as pointed out in Oughton and Lehr (2022).

1.3 Proposals

Recent developments have clarified essential requirements and practical aspects related to cell-free mMIMO networks, transforming them into more than a theoretical concept. This paves the way for more detailed evaluation frameworks for cell-free mMIMO network implementations, which are crucial for assessing the feasibility of cell-free mMIMO networks and their enablers. In light of this, the hypotheses of this thesis are examined in the studies outlined in the following subsections.

1.3.1 Fronthaul reliability analysis and protection schemes for cell-free mMIMO

The fronthaul network of a cell-free mMIMO system can be very complex due to the expected high density of deployed TRPs needed for a mMIMO operation with distributed antennas. Both industry and academia have proposed cell-free mMIMO with segmented fronthaul, which contains serial connections between TRPs in the fronthaul network, to reduce the fronthaul complexity and make the deployment more scalable and feasible. The hypothesis raised in this study is that segmented fronthaul may make cell-free mMIMO less feasible if there are no mechanisms to combat equipment failures at TRPs and fronthaul segments. In order to validate this hypothesis, this thesis proposes a reliability assessment methodology based on graph theory and MCMC simulation techniques, which can determine the system's SE for any instantaneous and cumulative failure configurations. Analyses are then performed in an indoor office scenario relevant to the industry to identify the failure impacts and possible protection strategies.

1.3.2 Techno-economic analysis for cell-free mMIMO

As the cell-free mMIMO technology progresses from a theoretical concept to a practical application, it is still unclear what is the most feasible way to perform its deployment. Since the technology's inception, there have been discussions in favor and against centralized or distributed processing implementations. Centralized processing offers higher interference cancellation performance, but comes with much higher computational costs. On the other hand, distributed processing has meager computational costs and is very flexible for cell-free mMIMO, as the system architecture is distributed. The hypothesis of this study is that, in order to determine the feasibility of different processing alternatives, a comprehensive technical and economic analysis is required, and that the doubts in the literature about which processing approach should be adopted persist because the economic feasibility of cell-free mMIMO networks has yet to be thoroughly analyzed. While some initial studies have been conducted, they have overlooked crucial operational aspects such as user demands, fronthaul limitations, and realistic hardware processing capacities and requirements. This hypothesis is validated in the thesis with the proposal of a comprehensive cost assessment methodology for cell-free mMIMO networks, which considers all aforementioned essential operational aspects. Under the methodology, analyses are performed in a dense urban scenario to answer in a clear way which situations make centralized or distributed processing implementations more feasible regarding demands, hardware, and cost variations.

1.4 Objectives

This thesis aims to develop comprehensive tools for analyzing the deployment feasibility of cell-free mMIMO networks. These tools should be able to represent and assess different deployment options and strategies, providing implementation guidelines that improve network feasibility.

In this context, this thesis seeks to offer robust methodologies and frameworks that can significantly contribute to the practical application and efficiency of cell-free mMIMO networks, considering scenarios of interest to society and industry.

The specific objectives of this thesis include:

1. Development of a reliability evaluation framework that can accurately simulate and assess the impacts of instantaneous and cumulative hardware failures in cell-free mMIMO systems.

2. Study of the reliability level of cell-free mMIMO networks to evaluate their resilience under different practical aspects, including serialization levels (the number of TRPs connected serially), pilot contamination, systems without ongoing maintenance supervision, and those with scheduled maintenance.

3. Development and evaluation of protection schemes that mitigate the effects of hardware failures in cell-free mMIMO networks, particularly focusing on segmented fronthaul configurations, both in non-integrated and integrated systems.

4. Development and implementation of a technical-economic assessment methodology for cell-free mMIMO networks that:

   - Consider fronthaul bandwidth and quantization limitations, TRP and CPU processing capacities, and adequate network dimensioning to support user demands
   - Predict the required number of active TRPs based on specific coverage or capacity constraints to support a given user rate or demands with and without minimum fairness requirements
   - Establish a method for calculating the fronthaul bit rate that allows minimal SE degradation, suitable for both distributed and centralized processing setups in cell-free mMIMO networks.
   - Determine the most feasible way to implement the processing in an urban scenario by comparing distributed and centralized functional split under three different popular high-performance linear scalable precoders: LP-MMSE, P-RZF, P-MMSE

The specific objectives 1 to 3 are related to the first hypothesis of the thesis, which is detailed in Subsection 1.3.1. Meanwhile, specific objective 4 and its sub-goals are associated with the second hypothesis, outlined in Subsection 1.3.2.

1.5 Contributions

The contributions of this thesis can be summarized as:

- A reliability framework for cell-free mMIMO networks capable of representing instantaneous and cumulative hardware failures over time.
- Fronthaul protection schemes designed to offset the effects of individual and cumulative hardware failures in cell-free mMIMO networks with segmented fronthaul.
- A techno-economic framework for cell-free mMIMO with a comprehensive methodology concerning essential operational parameters.
- A cost model for cell-free mMIMO networks, considering acquisition and installation of equipment and links, maintenance, floor space rent, and power consumption.
- Equipment distribution and resource allocation models for cell-free mMIMO networks
    - A TRP deployment model that can prioritize user demands or user fairness.
    - An adequate allocation of processing operations between CPU and TRPs for distributed and centralized processing functional splits.
    - A fronthaul bit rate model and suboptimal bit allocation methods based on a proposed maximum acceptable SE degradation due to quantization.
    - Models for hardware price and and energy consumption in TRPs and edge CPU.
- Insightful discussions on system reliability and cost-effectiveness of cell-free systems in indoor offices and dense urban scenarios.

1.6 Document organization

Beyond the introduction, this thesis is divided into five chapters, that are described as follows:

- Chapter 2 discusses the theoretical basis for the cell-free mMIMO networks analyzed in this thesis, describing a system model used in the chapters related to the proposals description and analysis.

- Chapter 3 discusses the theoretical basis for the understanding of the proposed reliability and economic frameworks described in the latter chapters, introducing graph theory, Continuous-Time Markov Chain (CTMC), and basic techno-economics concepts.

- Chapter 4 is dedicated to the first proposal of this thesis, which involves the reliability analysis and protection schemes for fronthaul in cell-free mMIMO systems. This chapter introduces a comprehensive framework designed to evaluate the reliability aspects. It proposes protection schemes to increase reliability in systems with separated fronthaul segments and TRPs, or when those equipment are integrated into just one device. Finally, it presents in-depth evaluations of indoor office scenarios to understand and enhance the robustness and efficiency of cell-free mMIMO networks.

- Chapter 5 addresses the second proposal of this thesis, which centers on the techno-economic analysis of cell-free mMIMO systems. It introduces a comprehensive cost methodology that accounts for user demands, with appropriate TRP distribution, as well as fronthaul and computational capacity requirements. It presents in-depth evaluations comparing the costs associated with centralized and distributed processing alternatives for a cell-free mMIMO network deployed at dense urban areas, aiming to determine the feasibility of each processing approach.

- Chapter 6 presents the final considerations concerning both proposals related to the thesis objectives, presenting avenues of relevant future works.

The thesis also includes an appendix covering technologies and architectures for cellular networks. While this topic is not directly related to the main proposals, it may still interest some readers by providing a broad overview of current mobile network solutions. This can enhance the reader's understanding of the context in which the main proposals are situated and how they can be expanded for future works.

2 Theoretical background and system model for cell-free mMIMO

This chapter provides the theoretical background necessary to comprehend the operation and implementation of Cell-Free mMIMO networks. It explores their architecture, properties, transport structure, transmission format, signal processing techniques, and the unique characteristics that differentiate them from traditional cellular networks. A mathematical system model for a Cell-Free mMIMO network with fully coherent transmission is presented. The system model is an input for the framework presented in Chapters 4 and 5, being used to calculate the SE achieved by the different TRP, UE, fronthaul and processing configurations. The content of this chapter is based prior art, the only exception is the distribution of the processing tasks between TRPs and CPUs concerning two functional splits for cell-free mMIMO, which are common in the literature for distributed and centralized processing implementation approaches. This distribution is a contribution of this thesis.

2.1 Cellular networks

A cellular network is a type of Radio Access Network (RAN) that provides voice and data services to mobile UE. It consists of cells with a particular non-overlapping coverage area containing one radio transceiver hub, called BS or TRP, that communicate with the UEs. The need for a cellular network arises because as the electromagnetic waves propagate from the transmitter, the signal energy spreads out. Consequently, less energy reaches a desired receiver as the distance increases. In this way, to guarantee a good Signal-to-Noise Ratio (SNR) anywhere on the network, the served area needs to be divided into cells that operate individually (Björnson; Hoydis; Sanguinetti, 2017).

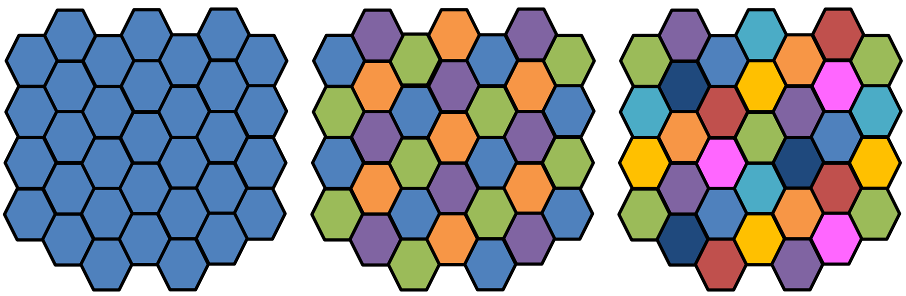
Mobile networks based on the cellular paradigm have undeniably revolutionized wireless communication, serving as the main driver in delivering wireless services over the last forty years. However, inherent to their design, these networks face a significant challenge: each cell operates independently, leading to potential interference issues, especially in areas where cells are in close proximity. In such scenarios, interference problems are exacerbated as the energy of interfering signals is less dissipated (Björnson; Hoydis; Sanguinetti, 2017). A critical interference case occurs at cell borders, where the desired signal energy is at its weakest while the interfering signal energy from adjacent cells reaches its peak. This interference phenomenon poses a considerable obstacle to achieving optimal network performance and user experience (Björnson; Jorswieck, 2013).

Over the years, intelligent code/time/frequency resource allocation has been used to minimize inter-cell interference. Essentially, channel resources are divided between cells and reused to maximize the distance between two cells using the same resource. Ideally, all adjacent cells will have different resources. Figure 2 presents examples of different levels of reuse factor, including 1, 4, and 9. In the first case, all cells use the same resources, resulting in very high interference between users, which will be particularly strong at the cell borders. In the second case, the resource is divided by 4, ensuring that no two adjacent cells share the same resource. This maintains SNR levels while reducing SINR, thereby increasing SE. However, the cost is that the total resource is divided by 4 (Björnson; Hoydis; Sanguinetti, 2017).

In the last case, represented at 2(c), the resource is divided by 9, achieving a two-cell distance between cells with the same resource, resulting in even lower interference. However, if it is assumed that the resource being divided is bandwidth, one could argue that this last situation provides less than half the bandwidth per cell compared to the one in 2(b) (Björnson; Hoydis; Sanguinetti, 2017).

Figure 2 – Examples of channel resource reuse patterns to combat interference in cellular networks.

(a) Reuse factor of 1. (b) Reuse factor of 4. (c) Reuse factor of 9.



Source: elaborated by the author.

If even higher interference cancellation is desired, more advanced schemes can be used to allocate unique resources to users at cell borders (Björnson; Jorswieck, 2013).

Despite these efforts, cellular networks will always be suboptimal in terms of capacity. Channel resources are sacrificed or not well-utilized due to interference. This fact is becoming a major problem in next-generation wireless communication systems since the demands are so high that the intercell distance is becoming small enough that interference levels can become unmanageable. (Demir; Björnson; Sanguinetti, 2021).

Readers interested in more aspects of current cellular networks, including their RAN architecture and the current Third Generation Partnership Project (3GPP) standards for these types of networks, Long Term Evolution (LTE) and New Radio (NR), are encouraged to read Appendix A. Despite this, please note that the content there is not necessary to understand any of the chapters of this thesis, and it should be viewed as additional relevant context and information.

2.2 Cell-free mMIMO networks

As discussed in Section 2.1, cellular networks will always be suboptimal in capacity due to inter-cell interference. Although high peak data rates are achievable when the UEs are close to the BS, diminished rates prevail across a large part of the coverage area, most notably at the cell borders. This variability significantly undermines the service reliability of a cellular network. A solution to this problem is to exploit the inter-cell interference through the collaborative coordination, processing, and transmission of user signals by multiple TRPs rather than just one. Such an approach substantially enhances the system-wide SINR through the entire coverage area and, by extension, the SE (Björnson; Jorswieck, 2013)(Interdonato et al., 2019).

Depending on the implementation and execution, collaborative coordination can virtually eliminate cell boundaries within mobile networks. This fact leads to the denomination of cell-free and represents a transformation of the operational paradigm in mobile networks. Despite this, it is necessary to acknowledge that not all forms of cooperation culminate in a cell-free architecture. For example, cellular LTE and NR frameworks can facilitate cooperation through their Coordinated MultiPoint (CoMP) with joint transmission functionalities. Despite this, the practical benefits of cooperation within 3GPP cellular systems have been limited. One of the main reasons for this outcome is the network-centric implementation adopted by the 3GPP standards (Interdonato et al., 2019).

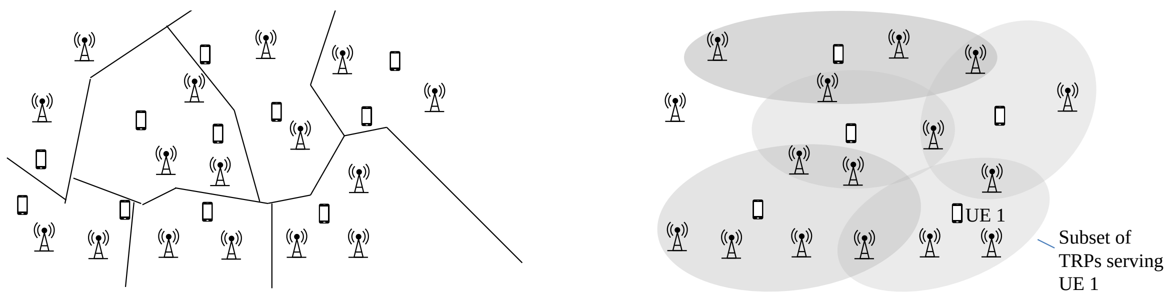
In the network-centric architecture, depicted in Figure 3(a), static clusters of neighboring BSs collaborate, functioning as a singular, expansive virtual cell. This methodology creates BS cooperation clusters that are inherently integrated into the operator's network infrastructure, facilitating channel acquisition, coordination, and synchronization within the cluster. Nonetheless, such a static cluster of BSs may yield suboptimal SE due to the uneven distribution of UEs and the presence of inter-cluster interference. Consequently, these clusters do not fully transcend the cellular paradigm, inheriting its limitations, which explains the minimal practical gains observed with CoMP in the context of LTE and NR (Björnson; Jorswieck, 2013)(Interdonato et al., 2019).

The problems associated with the network-centric method can be partly addressed by expanding cluster sizes, varying cluster configurations across frequency subcarriers, altering clusters dynamically, linking UEs within the cluster's range to a select number of BSs, and employing resource reuse strategies at cluster edges. However, these solutions are merely a *symptomatic treatment*. A more effective strategy is to implement CoMP with joint transmission using a user-centric approach. In this case, presented in Figure 3(b), the cluster of TRPs that serves or coordinates interference of a UE will be dynamically defined by the UE's needs. Consequently, each BS coordinates interference and serves with data two different sets of UEs, in such a way that the UEs served with the data are also in the interference coordination set. A general user-centric cooperation framework is DCC, which is defined as:

> "Any TRP j has channel estimates to users in $\mathcal{C}_j \subseteq \{1, ..., K\}$, where K is the number of users, while interference generated to users $i \in \mathcal{C}_j$ is negligible and can be treated as part of the Gaussian background noise. TRP j serves the users in $\mathcal{D}_j \subseteq \mathcal{C}_j$ with data."

Figure 3 – Possible BS cooperation implementations for network MIMO: (a) network-centric, and (b) user-centric.

(a) Network-centric. (b) User-centric.



Source: elaborated by the author.

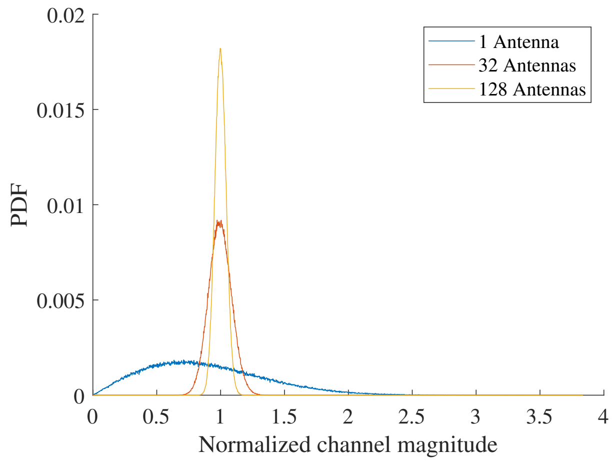
The user-centric cooperation facilitates a scenario where each UE can connect with a multitude of TRPs. In this case, each UE perceives the network as composed of a large number of distributed antennas in such a way that the mathematical framework developed for cellular mMIMO can be seamlessly applicable, allowing for the efficient cancellation of interference through simple linear processing techniques and hardware solutions. Consequently, cell-free mMIMO networks represent an evolution towards distributed, user-centric wireless non-cellular networks (in a traditional sense), integrating and enhancing the established methodologies of mMIMO for a new paradigm of connectivity, where ideally the number of TRPs is much higher than the number of UEs.(Björnson; Sanguinetti, 2020b)(Björnson; Hoydis; Sanguinetti, 2017).

2.2.1 Inherited properties from mMIMO and user-centric communication

As a merge of technologies, cell-free mMIMO inherits some essential properties important to understanding its inner workings and advantages concerning other types of communication systems, including channel hardening, favorable propagation, and macro-diversity.

The phenomenon known as channel hardening is inherited from mMIMO and makes a random fading channel behave as approximately deterministic. This occurs because, as the number of antennas used by TRPs to serve a UE increases, the law of large numbers of probability dictates that the overall channel behavior tends towards a less random pattern. More precisely, the aggregated characteristics of the channel are likely to stabilize around its mean value. This principle is visually represented in Figure 4, which plots the Probability Density Function (PDF) of the combined channel's magnitude, normalized by its mean. The illustration reveals a remarkable reduction in variation with 32 antennas compared to a single antenna setup. Moreover, increasing the antenna count to 128 further diminishes variation, signifying that the random channel behavior increasingly aligns with its mean as the number of antennas grows (Björnson; Hoydis; Sanguinetti, 2017).

Figure 4 – Illustration of an example explaining the channel property. Uncorrelated Rayleigh channels with unitary variance are assumed. The magnitude normalization is defined by the channel mean.



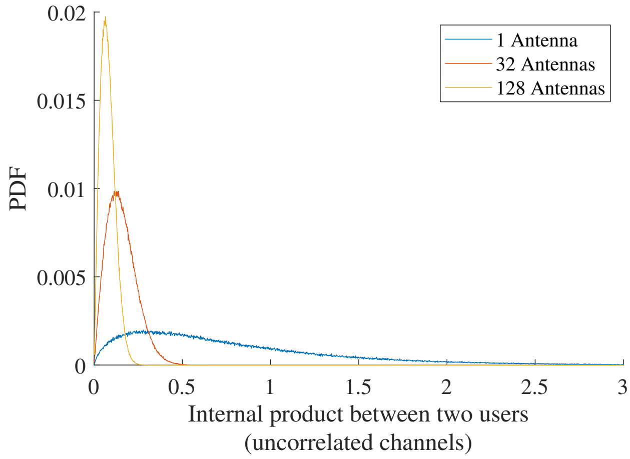
Source: elaborated by the author.

The channel hardening enables the UE to decode its information using channel statistics, which are significantly simpler to acquire than precise, real-time channel estimation. This approach eliminates the need to allocate substantial resources toward exact channel estimation in the downlink. Consequently, employing channel hardening-based decoding alongside Time Division Duplex (TDD) inside a coherence block simplifies and reduces the cost of UE processing. It ensures that only the TRP is required to perform channel estimation, facilitating a simpler open-loop MIMO operation, where there is no necessity for exchanging channel information between TRPs and UE, thereby avoiding unnecessary overhead (Demir; Björnson; Sanguinetti, 2021)(Björnson; Hoydis; Sanguinetti, 2017).

Despite these advantages, it is essential to note that the channel hardening exhibited by cell-free mMIMO systems may be substantially less pronounced than in traditional cellular configurations. As a result, certain UEs, especially those experiencing minimal channel hardening effects, may still require downlink channel estimation. This necessity has spurred the development of blind channel estimation techniques at the UEs, which aim to derive precise channel estimates directly from the transmitted user data, thus circumventing additional overhead (Attarifar; Abbasfar; Lozano, 2019)(Souza et al., 2022).

The property known as favorable propagation is also inherited from mMIMO and makes the channels of distinct users appear nearly orthogonal. The orthogonality between two random variables tends to increase with dimension. When the antenna count connecting TRPs to a UE is sufficiently high, the channels' dimensions for most users become large enough to achieve orthogonality. This effect is illustrated in Figure 5, which displays the PDF of the inner product between the channel vectors of two uncorrelated users across varying antenna counts. Mathematically, an inner product nearing zero indicates a higher degree of orthogonality between the channels. It is noticeable that with an increase in the number of antennas, the variance of the inner product diminishes, and the product value converges closer and closer to zero. Specifically, for configurations with 128 antennas, the PDF is predominantly concentrated in a range indicating an inner product less than 0.25, exemplifying the channels' progression towards orthogonality with higher antenna counts (Demir; Björnson; Sanguinetti, 2021)(Björnson; Hoydis; Sanguinetti, 2017).

Figure 5 – Illustration of an example explaining the favorable propagation property. Uncorrelated Rayleigh channels with unitary variance are assumed.



Source: elaborated by the author.

The favorable propagation simplifies wireless systems by enabling linear processing for interference cancellation, avoiding the need for complex non-linear processing. In simple terms, adding more antennas enhances interference cancellation *resolution* through linear processing. However, if two channels are highly correlated, merely increasing antennas may not suffice for interference cancellation, leading to low favorable propagation for the users involved. This situation usually happens at cellular systems when two users have the same direction (Björnson; Hoydis; Sanguinetti, 2017).

Nonetheless, due to the significant distances between cooperating TRPs in cell-free mMIMO, correlation is generally confined to antennas within each TRP, making UE channels largely independent. If the users are distant enough to not be served by very similar sets of TRPs, favorable propagation can be achieved even in systems with just one antenna per TRP. If more antennas per TRP are considered, cell-free mMIMO systems can achieve more significant favorable propagation than traditional cellular setups (Demir; Björnson; Sanguinetti, 2021).

Low favorable propagation occurs in cell-free mMIMO when two users are close to each other, as they will tend to be served by the same TRPs. In such cases, more antennas per TRP will be needed to attain favorable propagation (Demir; Björnson; Sanguinetti, 2021).

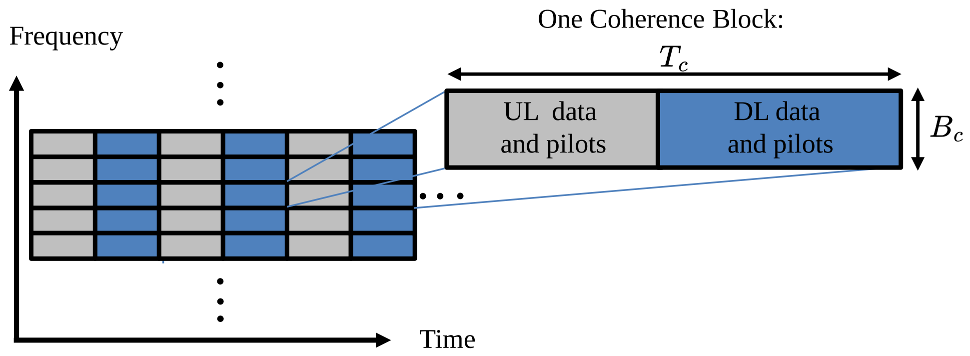
The property known as macro-diversity is inherited from user-centric communication and can mitigate fading effects by increasing the received signal strength and quality. In the downlink, each user will have multiple simultaneous information sources farther apart than the considered signal wavelength. In the uplink, the user sends its information to multiple receivers that work together, despite being at distance between themselves that is larger than the considered signal wavelength. In this way, the user can access more independent channels with a higher probability of at least some being strong. This advantage is not guaranteed in a cellular mMIMO scenario since the TRP antennas would be closer between themselves, and all channels components can be weak due to channel correlation (Gesbert et al., 2010).

2.3 Cell-free mMIMO system model

A cell-free mMIMO network consists of K mobile single antenna UEs and M_{tot} TRPs, each equipped with one or more antennas and arbitrarily distributed over the coverage area. Each TRP is connected through a fronthaul link to CPU, which is responsible for the coordination between multiple TRPs. Finally, the CPUs connect themselves and the network operator backbone through backhaul links. This structure allows the processing of channel estimation, precoding, and combining to be made at the CPU, in a centralized way, and locally at the TRPs, in a distributed way. The cost of opting for the latter is a more complex TRP and the exclusive utilization of more limited signal processing techniques (Demir; Björnson; Sanguinetti, 2021).

The system operates under multicarrier modulation under the TDD protocol inside a coherent time-frequency resource block, i.e., a number of sub-carriers and time samples over which the channel response can be approximated as constant and flat-fading. Then, $\tau_c = B_c T_c$ complex-valued samples will exist in each coherence block, where B_c and T_c are the coherence bandwidth and time, respectively. In this way, the total system frequency bandwidth and transmission time is divided into multiple coherence blocks, as represented in Figure 6. Each coherence block comprises uplink and downlink data, with a number of samples equal to τ_u and τ_d. For channel estimation, pilot signals are used in uplink and downlink with τ_{up} and τ_{dp} samples, and there will be $\tau_p = \tau_{up} + \tau_{dp}$ pilot samples per coherence block. However, it is common to use only uplink pilot signals due to channel hardening. Finally, all users utilize all coherence blocks and interference management is done by Space Division Multiple Access (SDMA) (Björnson; Sanguinetti, 2020b)(Demir; Björnson; Sanguinetti, 2021).

Figure 6 – The TDD multicarrier modulation scheme of a cell-free mMIMO network. The time-frequency plane is divided into coherence blocks in which each channel is time-invariant and frequency-flat.



Source: elaborated by the author.

The exact value of T_c or B_c depends on many physical factors and is hard to obtain. However, as a rule of rule-of-thumb, they can be approximated by:

$$T_c = \frac{\lambda}{4v} \tag{2.1}$$

$$B_c = \frac{1}{2T_d} \tag{2.2}$$

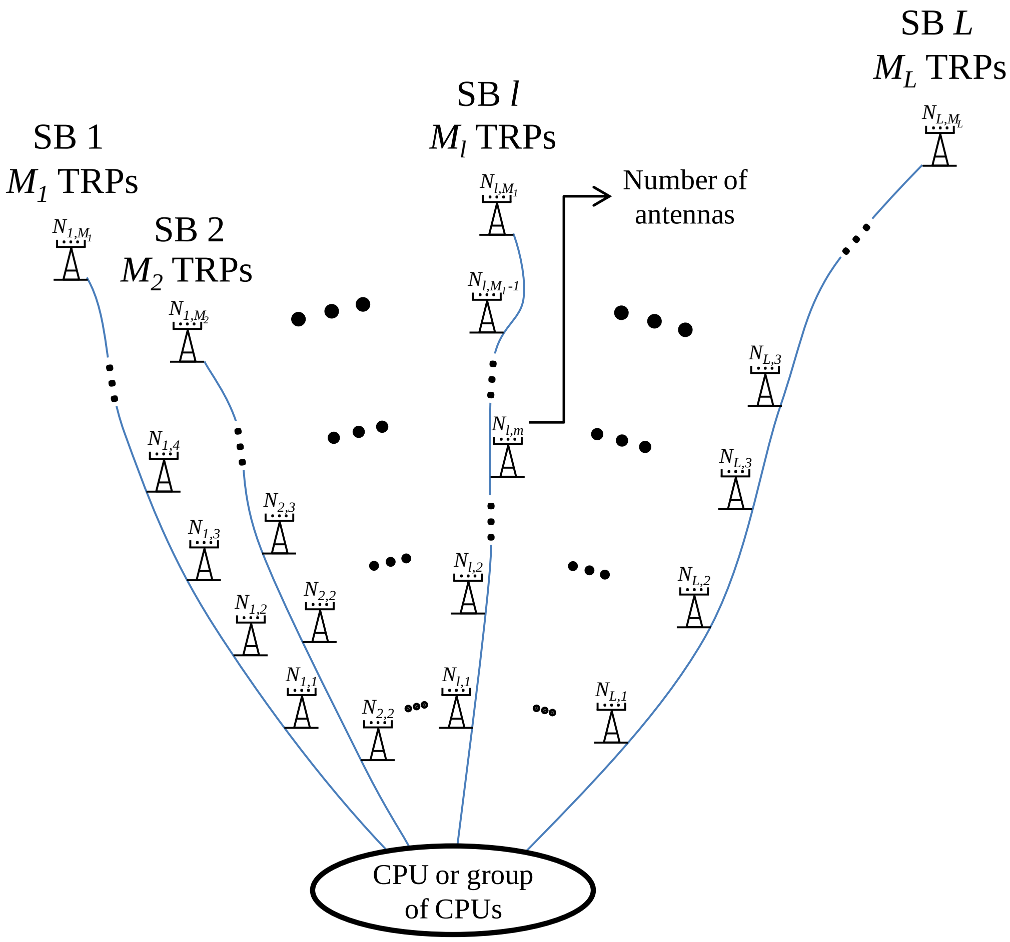
where λ, v, and T_d are the carrier wavelength, the user speed, and the delay spread (i.e., the time difference between the shortest and longest path), respectively (Björnson; Hoydis; Sanguinetti, 2017).

2.3.0.1 Cell-free mMIMO with segmented fronthaul

Traditionally, the underlying fronthaul design of cell-free mMIMO networks usually employs a star topology, with a separate link between each TRP and a CPU. However, a large number of TRPs may result in complex cabling for the fronthaul, which may be unscalable and cost-prohibitive for wide-area networks (Interdonato et al., 2019).

A solution to this problem is adopting a segmented fronthaul, where the TRPs are serially interconnected between themselves and a CPU in a Segmented Bus (SB) using a compute-and-forward communication architecture, as presented in Figure 7 (Interdonato et al., 2019). In this way, a cell-free mMIMO network will have L SBs of M_l serially interconnected TRPs, each having $N_{l,m}$ antennas and serving $K_{l,m}$ users. Then, the total number of antennas at the TRPs will be $N_{tot} = \sum_{l=1}^{L} \sum_{m=1}^{M_l} N_{l,m}$ and the total number of TRP will be $M_{tot} = \sum_{l=1}^{L} M_l$. Moreover, $K_{l,m} \leq K$ and the users served by each TRP are selected in a way to ensure K globally served users. This representation is compatible with both segmented and unsegmented fronthaul, in the latter case $L = M_{tot}$ and $M_l = 1$.

Figure 7 – Cell-free mMIMO network with segmented fronthaul L SBs of M_l serially connected TRPs with $N_{l,m}$ antennas serve K single-antenna users.



Source: elaborated by the author.

2.3.1 Channel models for cell-free mMIMO

Any wireless communication channel is modeled by large-scale fading and small-scale fading. The first does not vary rapidly and is generally related to distance-dependent path loss, shadowing, and the effects of transceiver hardware. The second does vary rapidly and mainly occurs due to multipath propagation, which practically always exists in terrestrial communication systems in such a way that the overlapping received signals can reinforce or cancel each other. Besides that, the channel itself may vary due to variations in the propagation environment.

Different models can describe the large-scale fading to different scenarios with different characteristics, like Radio Receiver (Rx)-Radio Transmitter (Tx) distance, Rx-Tx height difference, the operational environment, shadowing, and others. Some examples are the free-space, COST Hata, and the 3GPP propagation models.

The small-scale fading modeling will depend on the number of communication paths significant to the total overlapped received signal. The channel converges to a standard Gaussian distribution when this number is high due to the central limit theorem. This fact is the basis of the Rayleigh fading, a small-scale fading model applicable to a rich scattering environment. However, having LoS and NLoS channel components in practical systems is common, and the former can be stronger than the latter. In this case, Rician fading is a better small-scale model. It is a weighted sum of a deterministic channel variable related to LoS with a random one related to NLoS that is essentially a Rayleigh fading (Demir; Björnson; Sanguinetti, 2021) (Özdogan; Björnson; Larsson, 2019).

The channel is an essential input in the remainder equations of the system model described in this chapter. The analysis of the hypotheses in this thesis, presented in the latter chapters, mainly concentrates on correlated Rician fading. Nevertheless, uncorrelated Rayleigh fading is also employed for specific results. Nonetheless, the evaluation frameworks introduced in the latter chapters can be used with any models detailed in this subsection and possibly others in the literature.

2.3.1.1 Uncorrelated Rayleigh fading

The uncorrelated Rayleigh fading has elements $\mathbf{h}_{l,m,k} \in \mathbb{C}^{N_{l,m}\times 1}$ that are uncorrelated, and consequently independent with Rayleigh distributed magnitudes. It is a tractable model for rich scattering conditions, where the TRP antenna array is surrounded by many scattering objects compared to the number of antennas. This yields

$$\mathbf{h}_{l,m,k} \sim \mathcal{CN}(\mathbf{0}_{N_{l,m}}, \beta_{l,m,k}\mathbf{I}_{N_{l,m}}), \tag{2.3}$$

where $\beta_{l,m,k}$ describes the macroscopic large-scale fading between the TRP m on SB l to user k (Demir; Björnson; Sanguinetti, 2021).

2.3.1.2 Correlated Rayleigh fading

Practical channels generally have space-selective fading, i.e., they are spatially correlated. This happens because the antennas have non-uniform radiation patterns, and some spatial directions are more likely to carry strong signals due to the physical propagation environment. Therefore, a correlated Rayleigh fading is more realistic than an uncorrelated one, being given by

$$\mathbf{h}_{l,m,k} \sim \mathcal{CN}(\mathbf{0}_{N_{l,m}}, \mathbf{R}_{l,m,k}), \tag{2.4}$$

where $\mathbf{R}_{l,m,k} \in \mathbb{C}^{N_{l,m}}$ is the spatial correlation matrix that describes macroscopic propagation effects, including the antenna gains and radiation patterns at the transmitter and receiver. In such a way that the average channel gain (large-scale fading) between an TRP m on SB l to a user k is given by (Björnson; Hoydis; Sanguinetti, 2017)

$$\beta_{l,m,k} = \frac{1}{N_{l,m}} \mathrm{tr}(\mathbf{R}_{l,m,k}). \tag{2.5}$$

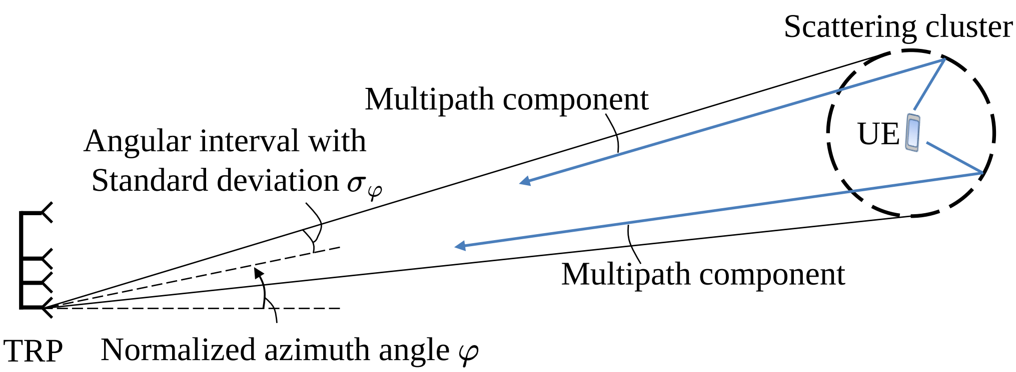
The spatial channel correlation of $\mathbf{h}_{l,m,k}$ is determined by the eigenstructure of $\mathbf{R}_{l,m,k}$, which will contain the statistical information on which spatial directions are more likely to contain strong or weak signal components. $\mathbf{R}_{l,m,k}$ will depend on the array geometry and the angular distribution of the multipath components (Demir; Björnson; Sanguinetti, 2021).

For the generally small TRPs (up to 16 antennas) envisioned for cell-free mMIMO networks, it is common to utilize a Uniform Linear Array (ULA). In this way, considering a ULA with half-wavelength antenna spacing and that all the multipath arrive from the far-field from scattering clusters, as presented in Figure 8, the elements of $\mathbf{R}_{l,m,k}$ can be computed as

$$[\mathbf{R}_{l,m,k}]_{i,j} = \beta \iint e^{\mathrm{j}\pi(i-j)\sin(\bar{\varphi})\cos(\bar{\theta})} f(\bar{\varphi}, \bar{\theta}) d\bar{\varphi} d\bar{\theta}, \tag{2.6}$$

where $\bar{\varphi}$ denotes the azimuth angle and $\bar{\theta}$ denotes the elevation angle of a multipath component. Moreover, $f(\bar{\varphi}, \bar{\theta})$ is the joint PDF of $\bar{\varphi}$ and $\bar{\theta}$, which is related to the Angular Standard Deviation (ASD) σ_ϕ and most of the times is supposed to be a jointly Gaussian distribution with mean centered at the far-field azimuth and elevation angles (Demir; Björnson; Sanguinetti, 2021).

Figure 8 – Illustration of the NLoS propagation under the local scattering model, where the scattering is localized around the UE. The figure only shows the azimuth plane and two of the many multipath components are indicated. The nominal angle φ, and the ASD σ_φ of the multipath components are key parameters to model the spatial correlation matrix.



Source: elaborated by the author.

2.3.1.3 Correlated Rician fading

An important aspect to consider in practical cell-free mMIMO systems is that the number of TRPs will be larger than in centralized mMIMO. This implies that the TRPs contributing significantly to the user signal will be closer, resulting in a stronger LoS channel component(Jin; Yue; Nguyen, 2021)(Polegre et al., 2020).

In this context, correlated Rayleigh fading may not be the most accurate channel representation for cell-free mMIMO networks since it is better suited for the NLoS channels, which are more likely to happen when the TRP is farther away from the user (Jin; Yue; Nguyen, 2021)(Polegre et al., 2020).

A more realistic channel modeling for the cell-free mMIMO networks is the independent correlated Rician fading, defined as

$$\mathbf{h}_{l,m,k} = \underbrace{\sqrt{\frac{\kappa_{l,m,k}\beta_{l,m,k}}{1+\kappa_{l,m,k}}}\mathbf{h}^{\mathrm{LoS}}_{l,m,k}}_{\overline{\mathbf{h}}_{l,m,k}} + \underbrace{\sqrt{\frac{\beta_{l,m,k}}{1+\kappa_{l,m,k}}}\mathbf{h}^{\mathrm{NLoS}}_{l,m,k}}_{\mathbf{g}_{l,m,k}}, \tag{2.7}$$

where the mean $\overline{\mathbf{h}}_{l,m,k} \in \mathbb{C}^{N_{l,m}\times 1}$ corresponds to the LoS component and $\mathbf{g}_{l,m,k} \sim \mathcal{CN}(\mathbf{0}_{N_{l,m}}, \mathbf{R}_{l,m,k}) \in \mathbb{C}^{N\times 1}$ represents the NLoS component that has a covariance matrix $\mathbf{R}_{l,m,k} = \mathbb{E}\{\mathbf{g}_{l,m,k}\mathbf{g}^{\mathrm{H}}_{l,m,k}\} \in \mathbb{C}^{N_{l,m}\times N_{l,m}}$ describing the spatial correlation (Özdogan; Björnson; Larsson, 2019). Moreover, $\kappa_{l,m,k}$ is the Rician factor, that weights the model according to the participation of the LoS and NLoS components to the channel(Ozdogan; Bjornson; Larsson, 2019).

Generally, the Rician factor is assumed to be a function of the distance between the TRP m in SB l and UE k ($d_{l,m,k}$), being calculated as $10^{1.3-0.003d_{l,m,k}}$. However, it is possible to evaluate the LoS probability according to the propagation scenario (Özdogan; Björnson; Larsson, 2019). In 3GPP propagation scenarios there are equations already defined for this parameter. Then, the Rician factor can be calculated as $p_{\mathrm{LoS}}(d_{l,m,k})/(1-p_{\mathrm{LoS}}(d_{l,m,k}))$, where $p_{\mathrm{LoS}(d_{l,m,k})}$ is the result of LoS probability function for a distance $d_{l,m,k}$ (Ozdogan; Bjornson; Larsson, 2019)(3GPP, 2020).

2.3.2 Scalability aspects for cell-free mMIMO

Cell-free mMIMO uses a large number of TRPs to serve numerous UEs, ideally with more TRPs than UEs. In this context, a large-scale network. Accordingly to (Björnson; Sanguinetti, 2020b) scalability in a cell-free mMIMO network is achieved when the computational complexity remains finite and manageable for the following processing tasks as the number of users and TRPs approaches infinity:

- Signal processing for channel estimation;

- Signal processing for data reception and transmission;
- Fronthaul signaling for data and Channel State Information (CSI) sharing;
- Power control optimization.

A cell-free mMIMO where all TRPs are connected to all users is not scalable. In this case, when the number of users grows to infinity, the number of channel estimates needed for each TRP also grows to infinity. A crucial step in the direction of scalability is to set a maximum amount of users per TRP in such a way that the UE selection at the TRP is user-centric based, like in the DCC framework, as defined in Section 2.2. In this way, each TRP will be connected to a limited number of users even when the total number of users grows to infinity (Björnson; Sanguinetti, 2020b).

Another essential step to attain scalability is restricting the number of TRPs to which a UE can connect. When the TRPs count grows to infinity, the frequency of precoding/combining computations also do, regardless of TRP user selection practices. While connecting a UE to a large cluster of TRPs is ideal, excessively large clusters may not yield proportional benefits. Distant TRPs contribute minimally to the user's signal, and the overhead from signaling and co-processing in vast TRP clusters can detrimentally affect network performance. A rule-of-thumb is to ensure users perceive a mMIMO experience, ideally with access to 32 or 64 antennas. However, studies have shown that even smaller TRP clusters can achieve SE comparable to a fully connected network (Freitas et al., 2023).

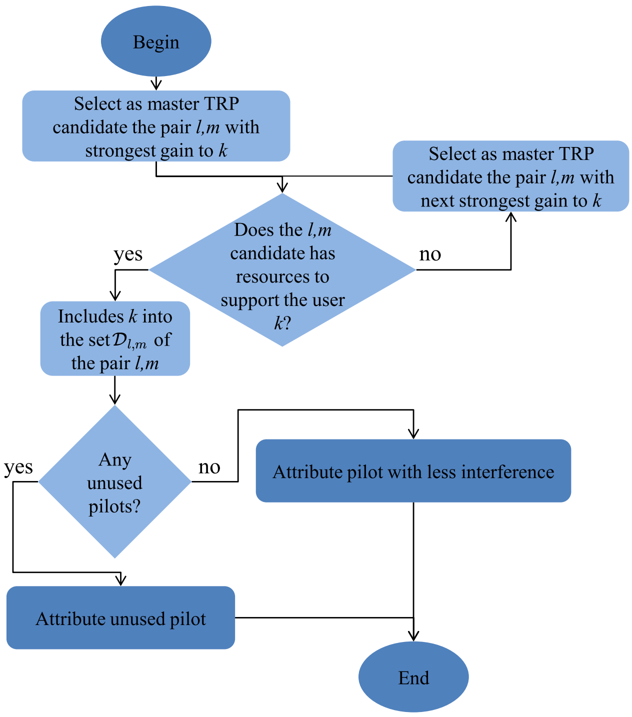
In DCC, the sets of users served with data ($\mathcal{D}_{l,m}$) and with coordination of interference ($\mathcal{C}_{l,m}$) of each TRP must be selected under the following conditions (Björnson; Jorswieck, 2013):

- The TRPs and objects in the propagation environment are static.
- The proximity between a user and an TRP is given by the average gain of the channel, taking into account the different transmission powers of the various TRPs.
- The channels between an TRP and the UEs that it seeks to serve or coordinate interference must be estimated.
- Each active UE must have a master TRP that guarantees its data services. This ensures no one is left without service, creating a natural hierarchy between the various BSs participating in a joint broadcast.
- The backhaul/fronthaul infrastructure must support joint transmission to a UE, enabling fast exchange of control signals and phase synchronization. Furthermore,

sufficient bandwidth is required to deliver the same data signal to all cooperating TRPs.

The master TRP selection for each user k follows the procedure depicted in Figure 9. This process creates initial sets $\mathcal{D}_{l,m}$ that only include users of which the TRPs are masters and assign pilot sequences seeking to minimize interference. When the user count exceeds the number of available pilots, interference estimation for a given pilot is performed by summing the channel gain over noise for users sharing that pilot in relation to the candidate master TRP. The pilot demonstrating the lowest interference level is then allocated to the user. The secondary user selection procedure for TRPs is shown in Figure 10, considering pilot availability and TRP hardware capacities, including maximum connection limits and required channel quality. Typically, the maximum number of connections at a TRP matches the pilot count, allowing all pilots to be utilized across TRPs. However, hardware constraints might reduce connections below the pilot count. The presented procedures are based on the framework presented in Björnson and Sanguinetti (2020b), which further details how the flowcharts of both master and secondary user selection can be made in a dynamic way compatible with DCC, including re-selection over time, TRP to TRP communication, and master status transfer.

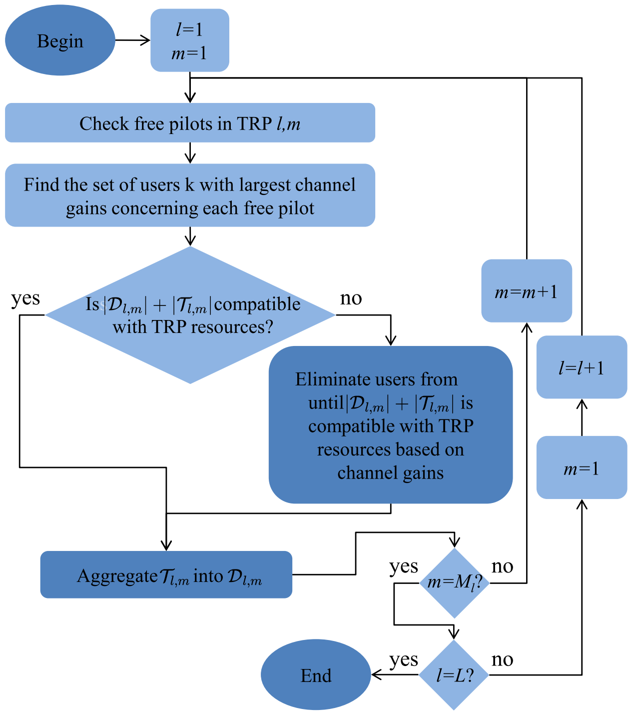
Figure 9 – Illustration of the master TRP selection procedure for a UE k. All users perform this procedure upon initial access to the network. The resources of a TRP can be understood as the minimum channel conditions and the supported number of UE connections.



Source: elaborated by the author.

Upon forming $\mathcal{D}_{l,m}$ for all $l \in 1, \cdots, L$ and $m \in 1, \cdots, M$, it is possible to establish the set of TRPs connected to each UE, $\mathcal{M}_k$, for each $k \in 1, \cdots, K$. The elements of the

Figure 10 – Illustration of the secondary user selection at all TRPs. This procedure is performed after all UEs have performed the their master TRP selection.



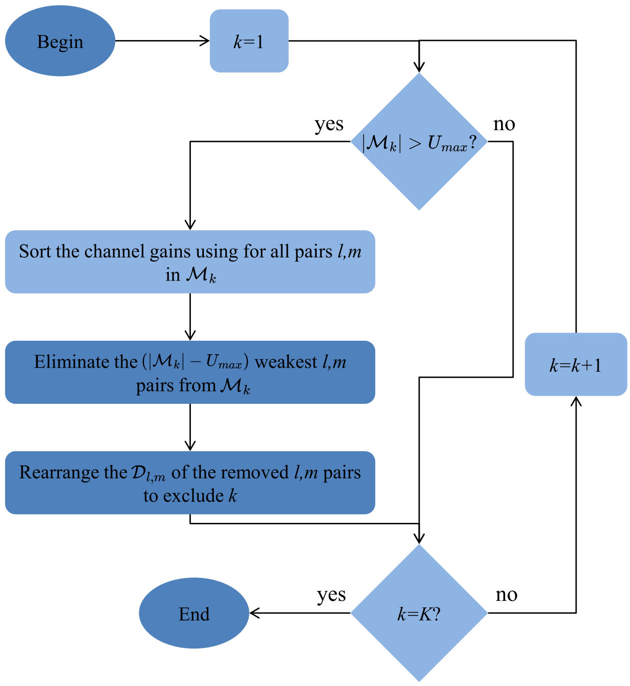
Source: elaborated by the author.

sets $\mathcal{M}_k$ are l, m pairs including k within their $\mathcal{D}_{l,m}$ set. To ensure scalability, both $\mathcal{D}_{l,m}$ and $\mathcal{M}_k$ are adjusted to ensure that the number of TRPs connections per UE does not exceed a predefined maximum, U_{max}, as detailed in the flowchart in Figure 11. Finally, with the sets formed, it is possible to formulate diagonal matrices that represent which TRPs a given UE k is connected to

$$\mathbf{D}_k = \begin{bmatrix} \mathbf{D}_{1,1,k} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & & & & & \vdots \\ 0 & & \mathbf{D}_{1,m,k} & & & & & & 0 \\ \vdots & & & \ddots & & & & & \vdots \\ 0 & & & & \mathbf{D}_{1,m,k} & & & & 0 \\ \vdots & & & & & \ddots & & & \vdots \\ 0 & & & & & & \mathbf{D}_{l,m,k} & & 0 \\ \vdots & & & & & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & \mathbf{D}_{L,M,k} \end{bmatrix}, \tag{2.8}$$

where $\mathbf{D}_{l,m,k}$ is equal to $\mathbf{I}_{N_{l,m}}$ if k belongs to $\mathcal{D}_{l,m}$ or $\mathbf{0}_{N_{l,m}}$, otherwise. This matrix formulation interfaces connection sets with the employed channel models. Essentially, if all TRPs were connected to all UEs, the aggregated channel of a UE k would be $\mathbf{h}_k = [\mathbf{h}_{1,1,k}^T, \cdots, \mathbf{h}_{1,M_1,k}^T, \mathbf{h}_{2,1,k}^T, \cdots, \mathbf{h}_{2,M_2,k}^T, \cdots, \mathbf{h}_{l,m,k}^T, \cdots, \mathbf{h}_{L,1,k}^T, \cdots, \mathbf{h}_{L,M_L,k}^T]^T$ and in the case of partial scalable connection, the aggregated channel is $\mathbf{D}_k\mathbf{h}_k$. It is important to point out that the fully connected network can be represented by making $\mathbf{D}_k$ an identity matrix (Björnson; Sanguinetti, 2020b).

Figure 11 – Illustration of UE connections adjustment for scalability with increased TRP count. This procedure is performed after master and secondary user selection.



Source: elaborated by the author.

2.3.3 Functional splits and fronthaul requirements for cell-free mMIMO

In terms of deployment architecture, a cell-free mMIMO network closely resembles a cellular Centralized Radio Access Network (C-RAN) network. Each cell site in a C-RAN contains one or more radio units performing radio functions. Then, all baseband signal processing is performed at the Baseband Unit (BBU) pool, which consist of multiple BBUs or an edge cloud computing server running multiple virtual BBUs. Readers interested in cellular radio access architectures are encouraged to read Section A.2 in Appendix A.

Considering the mentioned similarity, in a cell-free mMIMO network, the CPU corresponds to a centralized pool of BBUs, and the TRPs align with the Radio Units (RUs) dispersed throughout the coverage area. Then, in a similar fashion the C-RAN implementation, the cell-free setup allows for the flexible allocation of various digital processing tasks between the CPU and TRPs, optimizing fronthaul requirements, simplifying deployment, and enabling centralized resource management.

The bulk of digital signal processing in a cell-free mMIMO network heavily relies on two critical procedures: channel estimation and precoding/combining. These functions are pivotal in mitigating interference within the system and facilitating robust SDMA performance. Channel estimation involves accurately determining the characteristics of wireless channels between transmitters and receivers. Precoding techniques involve pre-shaping signals transmitted from multiple antennas to exploit channel behaviors. On the other hand, combining techniques involves merging signals received from multiple antennas to leverage diversity gains. Both combining and precoding techniques improve SE and mitigate interference among users through SDMA.

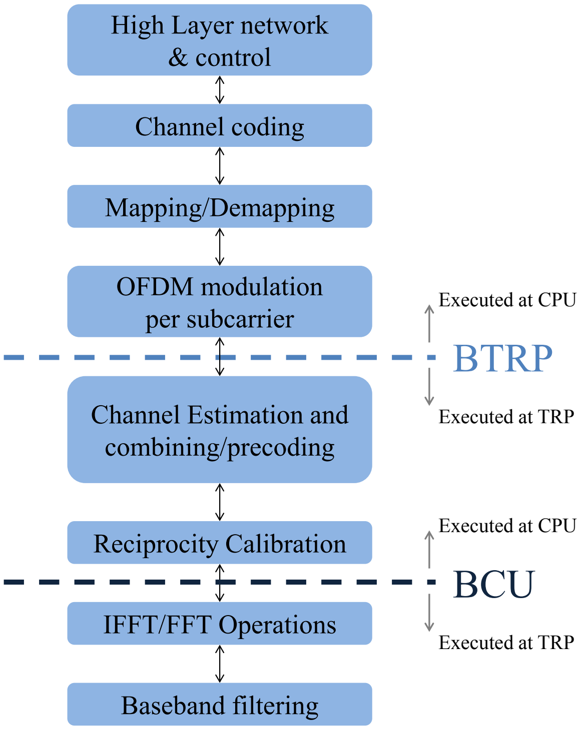
In the literature on cell-free mMIMO networks, the channel estimation and precoding/combining processing tasks happen at different types of hardware depending on two main functional splits: Baseband on Transmission-Reception Point (BTRP) and Baseband on Central Processing Unit (BCPU). The former split offloads these calculations to the TRPs, while the latter split centralizes them at the edge CPU (Femenias; Riera-Palou, 2020)(Björnson; Sanguinetti, 2020a). However, these are not the only tasks the system needs to perform. According to Demir et al. (2024), key processing tasks also include higher-layer control and network management, channel coding, mapping/demapping of signals, Orthogonal Frequency-Division Multiplexing (OFDM) modulation/demodulation per sub-carrier, reciprocity calibration, Fast Fourier Transform (FFT)/Inverse Fast Fourier Transform (IFFT) operations, and baseband filtering.

The complete task distribution among TRPs and edge CPUs in cell-free mMIMO networks, concerning BTRP and BCPU splits, was not fully detailed in the literature before the research associated with this thesis. Relevant works defining the splits, such as Femenias and Riera-Palou (2020) and Björnson and Sanguinetti (2020a), did not specify where tasks unrelated to channel estimation or combining/precoding were performed. Additionally, while Demir et al. (2024) considered all necessary processing tasks, it did not address any split equivalent to BCPU.

The task distribution proposed in this thesis for the mentioned functional splits, concerning TRPs and edge CPUs, is outlined in Figure 12, which includes all required processing tasks. It establishes that higher-layer control and network management, channel coding, mapping/demapping of signals, and OFDM modulation/demodulation per sub-carrier always occur at the CPU. Channel estimation, precoding, combining, and reciprocity calibration will happen at the CPU for BCPU and at the TRP for BTRP. Finally, IFFT/FFT operations always take place at the TRP.

In the context of cellular systems based on 3GPP standards, there exist established functional splits that do not directly correspond to those described in this section. This difference primarily arises from variations in processing execution and the sequence of operations, which are different from the one illustrated in Figure 12. Nonetheless, similarities can be observed with 3GPP split options, BCPU closely aligns with option 7.1, and BTRP fits within the spectrum of options 7.2 to 7.3. Thus, current 3GPP functional split processing options may be adaptable enough to support cell-free mMIMO. Readers

Figure 12 – Distribution of digital signal processing procedures in the CPU and TRPs for BCPU and BTRP splits.



Source: elaborated by the author.

interested in 3GPP functional split options for cellular networks are encouraged to read Section A.4 in Appendix A

In terms of resource requirements, the functional splits in literature for cell-free mMIMO also differ from 3GPP standards, mainly due to the type of network they are designed to support, especially concerning achieved fronthaul bitrate. In 3GPP splits a general notion is that the more centralized the processing, the higher the fronthaul bitrate achieved. Thus, the more centralized 3GPP 7.1 split achieves higher fronthaul bitrates than the more distributed 7.2 or 7.3 splits. However, when considering the traditional cell-free mMIMO splits from the literature, BCPU, which represents a more centralized approach, typically achieves lower fronthaul bitrates than BTRP.

This inversion occurs because both 3GPP 7.1 and BCPU fronthaul demands scales with the number of antennas in the TRP, whereas 7.2/7.3 and BTRP fronthaul demands scales with the number of independent beams or streams of data directed towards different directions or users. In cell-free mMIMO, this number of streams equals the number of users per TRP, whereas, in cellular 5G NR, it is usually limited to a maximum of 8 streams. Therefore, the number of streams is typically smaller than the number of antennas in 3GPP splits for cellular networks but larger than the number of antennas in cell-free mMIMO, as each TRP is expected to serve more users than the number of its antennas in the latter network.

2.3.3.1 Fronthaul distortion

The TRPs are connected to CPUs via a fronthaul with limited capacity. In this way, the antenna signals or precoded/combined UE data symbols are not sent through the fronthaul in an infinite precision fashion but in quantized versions. The fronthaul distortion will impact the functional splits of BTRP and BCPU in slightly different ways.

Figure 13 exemplifies the impacts of fronthaul distortion in a BTRP system. In the uplink, each antenna within a TRP receives both pilot and data samples, depicted by gray and brown blocks, respectively. The pilot signals are utilized to compute accurate channel estimates that are essential for determining the combiners $\mathbf{v}_k$ for each user. These combiners are then applied to the antenna signals to implement SDMA, enabling the extraction of distinct user data samples, represented by blocks of blue, red, and yellow colors. Up to this point, the system can be modeled with infinite precision, without bit width limitations for each sample [1].

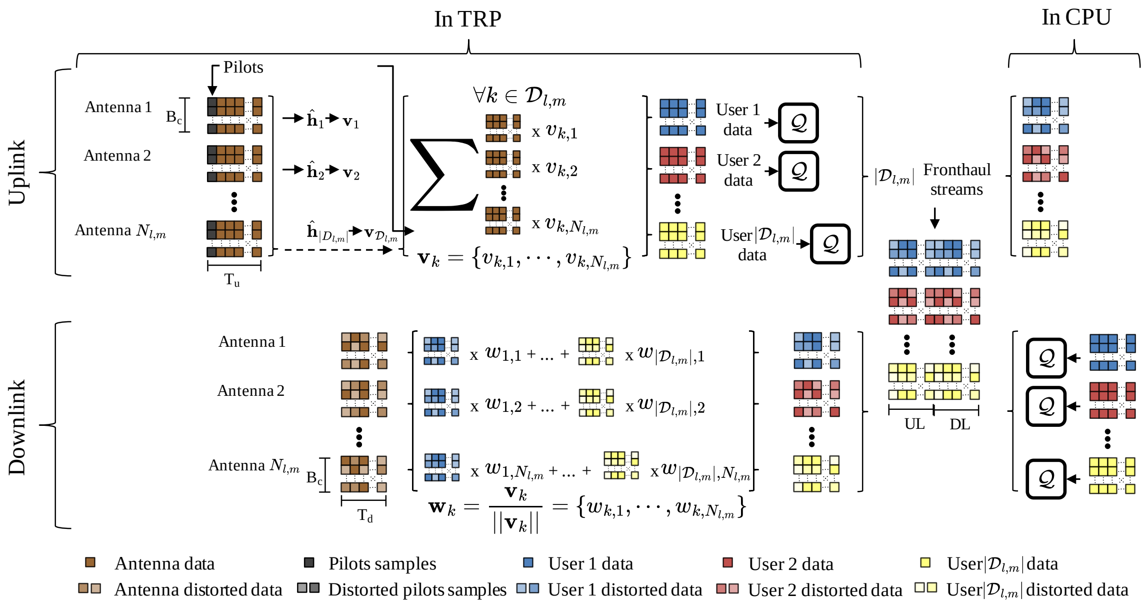
After obtaining the individual uplink user data samples, it becomes clear that the system depicted in Figure 13 undergoes a process of quantization, denoted by $\mathcal{Q}$. This process prepares the data for transmission over a limited fronthaul and is typically necessary due to the constrained fronthaul capacities, which restrict the number of bits available for representation for the samples. This restriction often affects the quality of transmitted data, as illustrated by variations in the color of user samples. This fact introduces quantization noise, and consequently, the CPU receives degraded data, increasing the likelihood of errors when decoding user information. Another important observation is that the TRP requires an amount of data streams proportional to its number of users for the fronthaul connection, and each stream transmits quantized uplink data of one user.

In the BTRP downlink operation, Figure 13 shows that the CPU modulates symbols for transmission at each sample, but must quantize them before fronthaul transmission. Consequently, the TRPs receive degraded downlink samples, which are then precoded and merged for transmission across all antennas to enable SDMA for interference cancellation. The precoder quality is high as it is derived from the combiner, calculated in the uplink phase within the TRP and unaffected by fronthaul constraints. Despite this, the input data for the precoding procedure is degraded due to fronthaul quantization. Then, the signal transmitted in each antenna of the TRP is also degraded, considering a system where no fronthaul losses would occur, and users have to deal with fronthaul quantization when decoding their downlink samples. Finally, an interesting behavior is that there are no exclusive fronthaul streams for downlink and uplink, the same stream transmits both,

[1] In practice, all pilots and data samples are subject to quantization in all network processing and transmission equipment due to the digital nature of the signal processing, i.e., quantization does not happen only on the fronthaul. However, the number of quantization bits utilized within TRP and CPU hardware typically surpasses that available in the fronthaul, where bandwidth limitation is a significant design aspect. Therefore, it is reasonable to consider only the quantization effects related to fronthaul transmission without loss of generality.

taking advantage of the TDD operation of cell-free mMIMO systems (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

Figure 13 – Illustration of the fronthaul quantization procedure for the BTRP functional split for uplink and downlink operations under a single TRP and CPU. The distortions in the samples of the coherence block caused by the quantization are emphasized throughout the channel estimation and combining/precoding procedures. It is shown that number of data stream in the fronthaul is proportional to the number of users served by the TRPs.

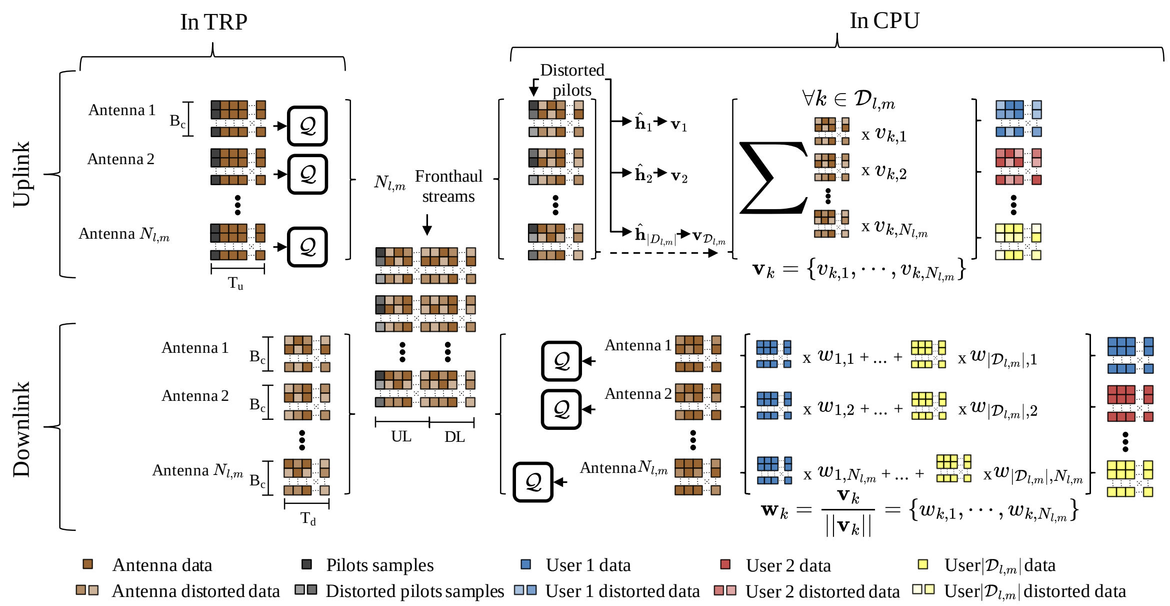


Source: elaborated by the author.

Figure 14 presents the impacts of fronthaul distortion in a BCPU system. Like the previous scenario, in the uplink operation, each antenna receives pilot and data samples, represented by gray and brown blocks, respectively. The difference is that the TRP does not perform combining. Instead, it quantizes the samples received at each antenna and transmits them directly to the CPU, which performs channel estimation and combining procedures. Consequently, the number of fronthaul streams is proportional to the number of antennas in the TRP. Subsequently, the CPU receives degraded antenna samples, illustrated by variations in color, including degraded pilot and data samples. The quantization noise in the pilots introduces errors that affect channel estimation accuracy, impacting the calculation of combiners. This fact can result in poor combining performance, making separating users' signals through SDMA difficult. After the combining procedure, separate user data samples are obtained, represented in red, blue, and yellow. However, the potential less accurate combiner and data sample quantization noise can cause unexpected variations in the received user data samples. The color variations in the user data sample blocks depict these variations, which can increase the likelihood of the CPU incorrectly decoding user samples and potentially limit system capacity.

In the BCPU downlink operation, Figure 14 shows that the CPU performs precoding and merges the user's signals for transmission across all antennas to enable SDMA. The precoding procedures have some errors because of the degraded channel estimation due to pilot quantization. Then, the antenna-precoded signals are quantized for transmission through the fronthaul, incurring additional quantization degradation that the users have to deal with fronthaul quantization when decoding their downlink samples. Finally, once again, there are no exclusive fronthaul streams for downlink and uplink due to TDD operation. (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

Figure 14 – Illustration of the fronthaul quantization procedure for the BTRP functional split for uplink and downlink operations under a single TRP and CPU. The distortions in the samples of the coherence block caused by the quantization are emphasized throughout the channel estimation and combining/precoding procedures. It is shown that number of data stream in the fronthaul is proportional to the number of antennas in the TRPs.



Source: elaborated by the author.

Based on the discussion in this subsection, it may seem that BTRP could offer better interference cancellation because it avoids distortion in channel estimates. However, this may not be the case, as in BCPU, the CPU can utilize channel estimates from multiple TRP to globally cancel interference, even if the received data is distorted.

2.3.3.2 Modeling distortions in generic quantized signals

The impacts of the errors associated with the quantization processes can be hard to model, mainly due to the nonlinear nature of the quantizers used to compress the signals to be transferred between TRPs and CPU. Despite this, a linear approximation of the quantized signal and error, called AQNM, can be obtained through the Bussgang decomposition (Bussgang, 1952). While not an exact approximation, this linear model for the quantization process has been extensively used in the mMIMO and cell-free mMIMO literature (Femenias; Riera-Palou, 2020)(Bashar et al., 2021)(Bashar et al., 2019)(Zhang et al., 2019).

Within the framework of the AQNM described in Femenias and Riera-Palou (2020), the relationship between any unquantized generic signal represented by an array, its quantized version, and the quantization error is described by the following

$$\hat{\mathbf{y}} = \mathcal{Q}_\theta(\mathbf{y}) = \mathbf{F}_\theta \mathbf{y} + \mathbf{q}_\theta, \tag{2.9}$$

where $\mathbf{y}$ represents the generic signal as an array with dimensions $n_y \times 1$, $\mathcal{Q}_\theta$ denotes the quantization function, $\mathbf{F}_\theta$ is the correlation matrix between the quantized signal $\hat{\mathbf{y}}$ and the original signal $\mathbf{y}$ with dimension $n_y \times n_y$, and $\mathbf{q}_\theta$ is the quantization error with dimensions $n_y \times 1$, modeled as a random variable that is uncorrelated with $\mathbf{y}$. The presented model also works when the input signal is a scalar or a matrix. In the former case, it is like an array with just one element, and in the latter case, each separate column of the matrix can be concatenated into a single-column array. After the estimated signal is obtained, the output array can be reshaped to the original matrix format. The matrix $\mathbf{F}_\theta$ correlates all pairs of elements of the random vectors $\mathbf{y}$ and $\hat{\mathbf{y}}$ while compensating auto-correlation aspects from the input signal $\mathbf{y}$, being calculated by

$$\mathbf{F}_\theta = \mathbb{E}[\hat{\mathbf{y}}\mathbf{y}^H]\left(\mathbb{E}[\mathbf{y}\mathbf{y}^H]\right)^{-1} = \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{R}_{\mathbf{y}\mathbf{y}}^{-1}, \tag{2.10}$$

where $\mathbf{R}_{\mathbf{y}\mathbf{y}}$ represents the auto-correlation matrix of generic original signal, and $\mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}$ denotes the cross-correlation matrix between the quantized signal and the original signal. If both $\mathbf{y}$ and $\hat{\mathbf{y}}$ are zero-mean variables, as it can be expected when using Quadrature Amplitude Modulation (QAM) signals, the presented correlation matrices can be considered to be covariance matrices (Björnson; Hoydis; Sanguinetti, 2017)(Gray; Davisson, 2004)(Pishro-Nik, 2014).

The error random vector in (2.9), $\mathbf{q}_\theta$, is calculated as

$$\mathbf{q}_\theta = \hat{\mathbf{y}} - \mathbf{F}_\theta \mathbf{y}, \tag{2.11}$$

and it will have an auto-covariance matrix given by

$$\mathrm{Cov}(\mathbf{q}_\theta, \mathbf{q}_\theta) = \mathbb{E}[(\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y})(\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y})^H] - \mathbb{E}[\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y}]\mathbb{E}[\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y}]^H, \tag{2.12}$$

$$\mathrm{Cov}(\mathbf{q}_\theta, \mathbf{q}_\theta) = \mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} - \mathbb{E}[\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y}]\mathbb{E}[\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y}]^H, \tag{2.13}$$

where $\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta}$ is the auto-correlation matrix of the quantization error. If both $\mathbf{y}$ and $\hat{\mathbf{y}}$ are zero-mean variables, $\mathbb{E}[\hat{\mathbf{y}} - \mathbf{F}_\theta\mathbf{y}]$ is also equal to zero and $\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta}$ becomes an auto-covariance matrix. The auto-correlation is given by

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbb{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^H - \hat{\mathbf{y}}\left(\mathbf{F}_\theta\mathbf{y}\right)^H - \left(\mathbf{F}_\theta\mathbf{y}\right)\hat{\mathbf{y}}^H + \mathbf{F}_\theta\mathbf{y}\left(\mathbf{F}_\theta\mathbf{y}\right)^H\right], \tag{2.14}$$

from the Hermitian transpose properties, it is known that $(\mathbf{M}_1\mathbf{M}_2)^H = \mathbf{M}_2^H\mathbf{M}_1^H$, then

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbb{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^H - \hat{\mathbf{y}}\mathbf{y}^H\mathbf{F}_\theta^H - \mathbf{F}_\theta\mathbf{y}\hat{\mathbf{y}}^H + \mathbf{F}_\theta\mathbf{y}\mathbf{y}^H\mathbf{F}_\theta^H\right], \tag{2.15}$$

since the sum of expectations is the expectation of the sums, and given the fact that $\mathbf{F}_\theta$ is not a random variable, it is possible to obtain

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbb{E}\left[\hat{\mathbf{y}}\hat{\mathbf{y}}^H\right] - \mathbb{E}\left[\hat{\mathbf{y}}\mathbf{y}^H\right]\mathbf{F}_\theta^H - \mathbf{F}_\theta\mathbb{E}\left[\mathbf{y}\hat{\mathbf{y}}^H\right] + \mathbf{F}_\theta\mathbb{E}\left[\mathbf{y}\mathbf{y}^H\right]\mathbf{F}_\theta^H, \tag{2.16}$$

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{F}_\theta^H - \mathbf{F}_\theta\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}} + \mathbf{F}_\theta\mathbf{R}_{\mathbf{y}\mathbf{y}}\mathbf{F}_\theta^H. \tag{2.17}$$

Finally, by replacing $\mathbf{F}_\theta$ using 2.10 the following is obtained

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{F}_\theta^H - \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{R}_{\mathbf{y}\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}} + \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{R}_{\mathbf{y}\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{y}\mathbf{y}}\mathbf{F}_\theta^H, \tag{2.18}$$

$$\mathbf{R}_{\mathbf{q}_\theta\mathbf{q}_\theta} = \mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} - \mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}\mathbf{R}_{\mathbf{y}\mathbf{y}}^{-1}\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}}, \tag{2.19}$$

where $\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}}$ is the cross-correlation matrix between the generic original signal and its quantized version, that will also be a covariance matrix if both $\mathbf{y}$ and $\hat{\mathbf{y}}$ are zero-mean variables (Gray; Davisson, 2004)(Pishro-Nik, 2014).

The calculation of the multiple correlation matrices $\mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}}$, $\mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}}$, and $\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}}$ can be done by knowing the quantization level, the input signal probability distribution, and the expected covariance between quantized and unquantized symbols. In Mezghani and Nossek (2012), closed-form expressions for these matrices are shown considering a Gaussian input and a simple 1-bit quantizer. However, it is very hard to obtain the values for the matrices for a general scalar quantizer, then approximations are typically necessary to obtain results for more complex multi bit quantizers (Femenias; Riera-Palou, 2020). The approach adopted in the system model presented at this chapter follows the methodology presented in Femenias and Riera-Palou (2020), assuming a zero-mean Gaussian distributions for the input and $\mathbf{q}_\theta$ under the following approximations

$$\mathbf{R}_{\hat{\mathbf{y}}\mathbf{y}} = \mathbf{R}_{\mathbf{y}\hat{\mathbf{y}}} \approx \alpha_\theta\mathbf{R}_{\mathbf{y}\mathbf{y}}, \tag{2.20}$$

$$\mathbf{R}_{\hat{\mathbf{y}}\hat{\mathbf{y}}} \approx (\alpha_\theta)^2\mathbf{R}_{\mathbf{y}\mathbf{y}} + \alpha_\theta(1-\alpha_\theta)\mathrm{diag}(\mathbf{R}_{\mathbf{y}\mathbf{y}}), \tag{2.21}$$

where α_θ is an empirical correlation coefficient between $\mathbf{y}$ and $\hat{\mathbf{y}}$, one that is an increasing function of bits per sample, and diag() is a function that extracts the diagonal elements of a square matrix to form a diagonal matrix.

This approach allows equation 2.9 to be approximated to

$$\hat{\mathbf{y}} = \mathcal{Q}_\theta(\mathbf{y}) \approx \alpha_\theta\mathbf{y} + \tilde{\mathbf{q}}_\theta, \tag{2.22}$$

where $\tilde{\mathbf{q}}_\theta$ is the Gaussian quantization error, modeled by $\mathcal{CN}\left(0, \alpha_\theta(1-\alpha_\theta)\mathrm{diag}(\mathbf{R}_{\mathbf{y}\mathbf{y}})\right)$. The approximation is valid because zero-mean is assumed for $\mathbf{y}$, and no correlation is assumed among the elements of $\tilde{\mathbf{q}}_\theta$ (Femenias; Riera-Palou, 2020).

The optimal design parameters of uniform and non-uniform quantizers may vary. Despite this, the signal distortion under the optimal configuration is tabulated by Max

Table 1 – Distortion factors of optimized uniform and non-uniform quantizers for Gaussian-distributed input signals with varying numbers of quantization bits per sample (Max, 1960).

Bits per sample	**Distortion rate**	**Correlation coefficient** (α_θ)
1	0.3634	0.6366
2	0.1175	0.8825
3	0.03454	0.96546
4	0.009497	0.990503
5	0.002499	0.997501
> 5	$1 - \alpha_\theta$	$1 - \frac{\pi\sqrt{3}}{2} 2^{-2(\text{bits})}$

(1960) for different quantization bits per sample, resulting in different α_θ as presented in Table 1.

This thesis uses the modelling presented in this subsection to obtain the fronthaul distortion and quantization noises in its entire system model and present analysis.

2.3.3.3 Fronthaul bitrate

The calculation of the fronthaul bit rate relies on the total number of coherence blocks across all available bandwidth within one second, the chosen functional split between TRPs and CPU, the interval between transmission of channel statistics to the CPU, the number of fronthaul transmitted samples in terms of real scalars, and the bit width to represent the samples (Demir; Björnson; Sanguinetti, 2021; Björnson; Sanguinetti, 2020b; Femenias; Riera-Palou, 2020).

Channel statistics typically change over time intervals significantly longer than the coherence time. Consequently, a large number of coherence blocks are transmitted through the fronthaul before there is a need to update these statistics. Statistics can often be inferred from the pilots and data within each coherence block, reducing the need to send statistical information explicitly. Consequently, for calculating fronthaul bit rate requirements, it is possible to rely solely on the pilot signals and instantaneous data samples contained within each coherence block without loss of generality. This approach effectively disregards the fronthaul traffic related to channel statistics without compromising accuracy or general applicability (Demir; Björnson; Sanguinetti, 2021).

The fronthaul bit rate in a cell-free system will also depend on the adopted fronthaul segmentation. In a distributed processing, when two serially TRPs connected serve the same user, they will exchange fronthaul data related to this user. Additionally, data pertaining to users not served by any two initial TRPs but by subsequent ones in the serial chain must also traverse the fronthaul link between aforementioned initial TRPs. In this way there will be a set of users served by a serial TRP chain $\mathcal{D}_l = \bigcup_{m=1}^{M} \mathcal{D}_{l,m}$ and a set of users in the chain starting from TRP m $\mathcal{D}_{l,\to m} = \bigcup_{m'=m}^{M} \mathcal{D}_{l,m'}$. Conversely, in a centralized processing model, there is no shared data between TRPs as each antenna receives an exclusive data stream from or to the CPU that performs combining/precoding procedures. Data intended for antennas in TRPs further down the serial connection line similarly has to be transmitted through the fronthaul connections of the initial TRPs, highlighting how the architecture influences fronthaul traffic and, consequently, the required bit rate.

When considering fronthaul segmentation, centralized precoders may be implemented in a distributed fashion by leveraging serial connections between TRPs. In such configurations, the fronthaul rate on the Serial Bus (SB) is not strictly correlated to the total antenna count across the TRPs within the SB. However, these setups do not fit the conventional BCPU and BTRP splits and are not explored in this thesis. Nonetheless, it is worth noting that these hybrid strategies can significantly lower fronthaul signaling requirements when deploying centralized precoders in cell-free mMIMO networks with fronthaul segmentation (Shaik; Björnson; Larsson, 2020; Miretti; Björnson; Gesbert, 2021).

In this context, the fronthaul bit rate to support an l, m TRP under the BTRP and BCPU split implementations at a given time t can be calculated by

$$F_{l,m,t} = \begin{cases} 2B\left(1 - \frac{\tau_p}{\tau_c}\right) \sum_{k \in \mathcal{D}_{l,\to m,t}} b^{\text{data}}_{l,k}, & \text{for BTRP} \\ \sum_{m'=m}^{M_l} 2N_{l,m'} B\left[\left(1 - \frac{\tau_p}{\tau_c}\right) b^{\text{data}}_{l,m'} + \frac{\tau_p}{\tau_c} b^{\text{pil}}_{l,m'}\right], & \text{for BCPU} \end{cases}, \tag{2.23}$$

where B is the total available bandwidth, $\mathcal{D}_{l,\to m,t}$ is $\mathcal{D}_{l,\to m}$ at time t, and $b^{\text{data}}_{l,k}$ is the bit width for the data symbols inside the coherence block between TRPs in the SB l and UE k in the BTRP implementation (Demir; Björnson; Sanguinetti, 2021; Femenias; Riera-Palou, 2020). Moreover, for the BCPU implementation, $b^{\text{data}}_{l,m'}$ is the bit width for the data samples of the coherence block in all antennas of TRP l, m', and $b^{\text{pil}}_{l,m'}$ is the bit width of pilot samples for channel estimation. The latter is applied only to τ_p samples of the coherence block (Demir; Björnson; Sanguinetti, 2021; Femenias; Riera-Palou, 2020). Different bit widths for data and pilots arise because a higher precision in channel estimation samples is usually necessary, implying large bit widths for pilots (Björnson; Sanguinetti, 2020a).

2.3.3.4 Computational complexity in hardware

The distinct tasks undertaken by TRPs and CPUs in BCPU and BTRP splits result in different computational requirements on these devices, which impact their energy consumption and costs. A way to model this required computational complexity is to quantify the amount of Giga Operations Per Second (GOPS) that the processing tasks in each device will require. However, determining the exact GOPS presents challenges, as the task complexities vary between the splits and scales differently with factors like bandwidth, number of users, system SE, and others.

Notably, the proposed division of processing functions in Figure 12 results on commonly executed on the CPU across both BCPU and BTRP splits. The computational complexity of these tasks is known for LTE cellular systems. In this context, since their execution in a cell-free context is mostly similar, these known GOPS values can be used as reference together with the resilient scaling model proposed in Debaillie, Desset and Louagie (2015) to estimate the CPU requirements that are common in BCPU and BTRP splits. Then, for a single CPU implementation, this thesis proposes that the processing complexity for CPU common operations between the two considered splits is given by

$$\begin{aligned} \mathrm{GOPS}_{t,R}^{\mathrm{CPUcommon}} = \gamma_{\mathrm{HLnt}}\,\mathrm{GOPS}_{\mathrm{HLnt}} + \gamma_{\mathrm{HLct}}\,\mathrm{GOPS}_{\mathrm{HLct}} + \gamma_{\mathrm{Chcd}} \\ \mathrm{GOPS}_{\mathrm{Chcd}} + \gamma_{\mathrm{MpDp}}\,\mathrm{GOPS}_{\mathrm{MpDp}} + \gamma_{\mathrm{OFDM}}\,\mathrm{GOPS}_{\mathrm{OFDM}} \end{aligned}, \tag{2.24}$$

were $\mathrm{GOPS}_{\mathrm{HLnt}}$, $\mathrm{GOPS}_{\mathrm{HLct}}$, $\mathrm{GOPS}_{\mathrm{Chcd}}$, $\mathrm{GOPS}_{\mathrm{MpDp}}$, $\mathrm{GOPS}_{\mathrm{OFDM}}$ are the reference values of GOPS for higher-layer network, higher-layer control, channel coding, layer mapping and demapping, and OFDM modulation, respectively. Moreover, γ_{HLnt}, γ_{HLct}, γ_{Chcd}, γ_{MpDp}, γ_{OFDM} are the scaling factors for the same tasks, respectively. Table 2 details the calculation of the different scaling factors, the variables B_{base} and $\mathrm{SE}_{\mathrm{base}}$ represent the bandwidth and SE of the reference GOPS value. In contrast, B and $\mathrm{SE}_{t,R}$ represent the adopted bandwidth and SE for an expected UE rate R, respectively.

Table 2 – GOPS scaling parameters calculation for common CPU operations in BCPU and BTRP (Debaillie; Desset; Louagie, 2015).

Scaling factor	**Calculation**
γ_{HLnt}	$\left(\frac{B}{B_{\mathrm{base}}}\right)^{1}\left(\frac{\mathrm{SE}_{t,R}}{\mathrm{SE}_{\mathrm{base}}}\right)^{1}$
γ_{HLct}	$(N_{tot})^{0.5}K^{0.2}$
γ_{Chcd}	$\left(\frac{B}{B_{\mathrm{base}}}\right)^{1}\left(\frac{\mathrm{SE}_{t,R}}{\mathrm{SE}_{\mathrm{base}}}\right)^{1}K^{1}$
γ_{MpDp}	$\left(\frac{B}{B_{\mathrm{base}}}\right)^{1}\left(\frac{\mathrm{SE}_{t,R}}{\mathrm{SE}_{\mathrm{base}}}\right)^{1.5}K^{1}$
γ_{OFDM}	$\left(\frac{B}{B_{\mathrm{base}}}\right)^{1}K^{1}$

Analyzing Table 2, it becomes evident concerning Debaillie, Desset and Louagie (2015) that the variable representing the number of streams has been substituted with the number of users for most scaling factors. This adjustment happens mainly due to each user being served with the full bandwidth using SDMA in the modeled cell-free network, making the number of spatial streams directly equivalent to the user count. Another notable difference from Debaillie, Desset and Louagie (2015) is observed in the OFDM scaling factor, which is proportional to the number of users instead of the number of antennas. This alteration arises from the fact that OFDM modulation per subcarrier occurs on a per-user basis in the presented cell-free splits instead of a per-antenna approach from Debaillie, Desset and Louagie (2015), in such a way that the modulated symbols transmitted to each antenna are derived through the precoding process. Lastly, the parameter for load has been omitted from all scaling factors since it is assumed to be always at 100%. Consequently, any exponents in Debaillie, Desset and Louagie (2015) related to load will not modify the scaling factors.

For an one CPU scenario, the proposed number of GOPS to be executed at the CPU for an expected UE rate R at the UE load of the time t can be calculated as

$$\mathrm{GOPS}_{t,R}^{\mathrm{CPU}} = \begin{cases} \mathrm{GOPS}_{t,R}^{\mathrm{CPUcommon}} + \mathrm{GOPS}_{t,R}^{\mathrm{BCPU}}, & \text{for BCPU} \\ \mathrm{GOPS}_{t,R}^{\mathrm{CPUcommon}}, & \text{for BTRP} \end{cases}, \tag{2.25}$$

where $\mathrm{GOPS}_{t,R}^{BCPU}$ is the number of GOPS of the specific CPU operations of the BCPU split for an expected UE rate R at the UE load of the time t, calculated as

$$\begin{aligned} \mathrm{GOPS}_{t,R}^{\mathrm{BCPU}} = & \underbrace{\frac{8N_{sc}CC_{\mathrm{all},t}^{\mathrm{est}}}{T_s 10^9 \tau_c}}_{\text{Channel estimation}} + \underbrace{\frac{8N_{sc}CC_{\mathrm{all},t}^{\mathrm{comb}}}{T_s 10^9 \tau_c}}_{\text{Precoding computation}} \\ & + \underbrace{\frac{8N_{sc}N \sum_{l=1}^{L} \sum_{l=1}^{L} \sum_{m=1}^{M_l} |\mathcal{D}_{l,m,t}|}{T_s 10^9 \tau_c}}_{\text{Reciprocity calibration}} + \underbrace{\frac{8N_{sc}N\left(\tau_c - \tau_p\right) \sum_{l=1}^{L} \sum_{m=1}^{M_l} |\mathcal{D}_{l,m,t}|}{T_s 10^9 \tau_c}}_{\text{Precoding}}, \end{aligned} \tag{2.26}$$

where N_{sc} is the number of subcarriers, T_s is the OFDM symbol duration, and $\mathcal{D}_{l,m,t}$ is $\mathcal{D}_{l,m}$ at time t. Moreover, $CC_{\mathrm{all},t}^{\mathrm{est}}$ and $CC_{\mathrm{all},t}^{\mathrm{comb}}$ denote the required number of complex multiplications and divisions in the CPUs to perform channel estimation and generate the precoding vectors for all active UEs at time t. These variables are computed by the equations in Table 3 at Section 2.3.4.4. The term $\frac{8}{(10^9 T_s)}$ converts the number of complex multiplications to GOPS. Additionally, reciprocity calibration is an one-time operation per coherence block. Thus, it is divided by τ_c. Finally, the precoder is exclusively applied to data samples, and as such, it is scaled by $(\tau_c - \tau_p)/\tau_c$ (Demir et al., 2024).

Multiple CPUs scenarios will differ in the sense that not all CPUs serve all the users, in such a way that the final CPU GOPS will be lower or equal to the single CPU case but the equation will have the same format. the derivation of the required complexity in this case is leaved as a future work and all analysis of thesis proposals concerning CPU required computational capacity will focus on a single CPU case.

In scenarios involving multiple CPUs, each CPU may not serve all users, which means that the overall CPU GOPS could be lower or equal to that in the single CPU scenario. However, the format of the equation remains the same. The derivation of the required complexity in multi-CPU cases is left as future work. Therefore, all analyses in this thesis regarding the computational capacity required by CPUs will focus solely on a single CPU scenario. Despite this limitation, this assumption has been deemed reasonable for the studied scenarios that involve analyzing computational capacities.

In both BCPU and BTRP cases, baseband filtering and IFFT/FFT operations are executed at the TRP. The proposed GOPS of these common operations for an expected

UE rate R at the UE load of the time t can be calculated as

$$\mathrm{GOPS}_{t,R}^{\mathrm{TRPcommon}} = \underbrace{\frac{8N_{DFT}\log_2(N_{DFT})}{T_s 10^9}}_{\text{FFT/IFFT}} + \underbrace{\frac{40Nf_s}{10^9}}_{\text{Baseband Filter}}, \tag{2.27}$$

where N_{DFT} represents the dimension of the Discrete Fourier Transform (DFT), and f_s is the sampling frequency. Moreover, The term $40Nf_s/10^9$ denotes the GOPS for a filter with ten taps in a polyphase filtering implementation (Demir et al., 2024).

The proposed number of GOPS to be executed at the TRP m in SB l for an expected UE rate R at the UE load of the time t can be calculated as

$$\mathrm{GOPS}_{t,R}^{\mathrm{TRP}_{l,m}} = \begin{cases} \mathrm{GOPS}_{t,R}^{\mathrm{TRPcommon}}, & \text{for BCPU} \\ \mathrm{GOPS}_{t,R}^{\mathrm{TRPcommon}} + \mathrm{GOPS}_{t,R}^{\mathrm{BTRP}_{l,m}}, & \text{for BTRP} \end{cases}, \tag{2.28}$$

where $\mathrm{GOPS}_{t,R}^{\mathrm{BTRP}_{l,m}}$ is the number of GOPS of the specific TRP l operations of the BTRP split for an expected UE rate R at the UE load of the time t, calculated as

$$\begin{aligned} \mathrm{GOPS}_{t,R}^{\mathrm{BTRP}_{l,m}} = & \underbrace{\frac{8N_{sc}CC_{l,m,t}^{\mathrm{est}}}{T_s 10^9 \tau_c}}_{\text{Channel estimation}} + \underbrace{\frac{8N_{sc}CC_{l,m,t}^{\mathrm{comb}}}{T_s 10^9 \tau_c}}_{\text{Combining computation}} \\ & + \underbrace{\frac{8N_{sc}N\,|\mathcal{D}_{l,m,t}|}{T_s 10^9 \tau_c}}_{\text{Reciprocity calibration}} + \underbrace{\frac{8N_{sc}N\,(\tau_c - \tau_p)\,|\mathcal{D}_{l,m,t}|}{T_s 10^9 \tau_c}}_{\text{Precoding}}, \end{aligned} \tag{2.29}$$

where $CC_{l,m,t}^{\mathrm{est}}$ and $CC_{l,m,t}^{\mathrm{comb}}$ denote the number of complex multiplications and divisions that the TRP m on SB l needs to perform channel estimation and generate the combining vectors for all active UEs at time t. Moreover, $CC_{l,m,t}^{\mathrm{est}}$ and $CC_{l,m,t}^{\mathrm{comb}}$ are computed as in Table 3 at Section 2.3.4.4.

2.3.4 Signal model for cell-free mMIMO

2.3.4.1 Pilot transmission and channel estimation

The system has τ_{up} orthogonal uplink pilots, which are used for uplink channel estimation. Ideally, each user k has an exclusive uplink pilot sequence to benefit from the pairwise orthogonality between the user's pilot signals to cancel interference in such a way that $\tau_{up} = K$. However, pilot sequences are a limited resource since the condition $\tau_p < \tau_c$ needs to be attended. Besides that, the utilization of very long pilot sequences may be computationally complex and consequently unfeasible in hardware. In this context, different UEs may be assigned to the same pilot sequence, resulting in $K > \tau_{up}$. This phenomenon is called pilot contamination and degrades the estimation quality and generates downlink coherent interference (Demir; Björnson; Sanguinetti, 2021).

For mathematical representation each pilot p is denoted by $\boldsymbol{\phi}_p \in \mathbb{C}^{\tau_{up}\times 1}$ and satisfy the following conditions: $||\boldsymbol{\phi}_p||^2 = 1$ and $\boldsymbol{\phi}_p^H \boldsymbol{\phi}_{p'} = 0$, $\forall p' \neq p$. Additionally, the index of a pilot associated with a UE k is denoted by $t_k \in \{1, \cdots, \tau_p\}$, and $\mathcal{P}_k$ denotes the sets of UEs that use the same pilot as the user k. From these considerations, the received pilot signals at the TRP m at SB l can be obtained as

$$\mathbf{Y}_{l,m}^{\text{pilot}} = \sum_{i=1}^{K} \sqrt{\tau_{up}\eta_k}\mathbf{h}_{l,m,k}\boldsymbol{\phi}_{t_i}^{\text{T}} + \mathbf{N}_{l,m}, \tag{2.30}$$

where η_k represents the uplink transmit power of UE k, and $\mathbf{N}_{l,m} \in \mathbb{C}^{N\times\tau_{up}}$ is the receiver noise with i.i.d. elements distributed as $\mathcal{CN}(0, \sigma_{ul}^2)$ (Demir; Björnson; Sanguinetti, 2021). In the BCPU split, the channel estimation is performed at the CPU using fronthaul-quantized pilot samples, which by using the AQNM approximation presented in subsection 2.3.3.1 are given by

$$\tilde{\mathbf{Y}}_{l,m}^{\text{pilot}} = \mathcal{Q}(\mathbf{Y}_{l,m}^{\text{pilot}}) = \tilde{\alpha}_{l,m}\mathbf{Y}_{l,m}^{\text{pilot}} + \check{\mathbf{N}}_{\mathbf{Y}_{l,m}}, \tag{2.31}$$

where $\tilde{\alpha}_{l,m}$ is the linear correlation coefficient between unquantized and quantized pilot signal versions and $\check{\mathbf{N}}_{\mathbf{Y}_{l,m}} \in \mathbb{C}^{N\times\tau_{up}}$ is the additive quantization noise, which is uncorrelated with $\mathbf{Y}_{l,m}^{\text{pilot}}$ (Zhang; Zhang; Ai, 2020)(Femenias; Riera-Palou, 2020).

The channel estimation procedure of a specific UE k projects the received pilots into the normalized pilot associated with k. This yields in $\mathbf{y}_{l,m,t_k} = \mathbf{Y}_{l,m}^{\text{pilot}}\boldsymbol{\phi}_{t_k}/\tau_{up}$ for BTRP and $\mathbf{y}_{l,m,t_k} = \tilde{\mathbf{Y}}_{l,m}^{\text{pilot}}\boldsymbol{\phi}_{t_k}/\tau_{up}$ for BCPU. In this way, the project pilots equivalent to

$$\mathbf{y}_{l,m,t_k}^{\text{pilot}} = \begin{cases} \sqrt{\tau_{up}\eta_k}\mathbf{h}_{l,m,k} + \sum_{i\in\mathcal{P}_k}\sqrt{\tau_{up}\eta_i}\mathbf{h}_{l,m,i} + \mathbf{n}_{l,m,t_k}, \text{ for BTRP} \\ \tilde{\alpha}_{l,m}\left(\sqrt{\tau_{up}\eta_k}\mathbf{h}_{l,m,k} + \sum_{i\in\mathcal{P}_k}\sqrt{\tau_{up}\eta_i}\mathbf{h}_{l,m,i} + \mathbf{n}_{l,m,t_k}\right) + \check{\mathbf{n}}_{l,m,t_k}, \text{ for BCPU} \end{cases}, \tag{2.32}$$

where $\check{\mathbf{n}}_{l,m,t_k} \sim \mathcal{CN}(0, \tilde{\alpha}_{l,m}(1-\tilde{\alpha}_{l,m})\tilde{\mathbf{Y}}_{l,m}\tilde{\mathbf{Y}}_{l,m}^H)$ is the additive quantization noise for the pilot of the UE k in TRP m at SB l, and $\mathbf{n}_{l,m,t_k}$ is additive thermal noise for the pilot of the UE k received at the antennas of TRP m at SB l(Zhang; Zhang; Ai, 2020)(Femenias; Riera-Palou, 2020).

There are several channel estimation methods with different degrees of complexity. The MMSE estimator has higher precision but is relatively complex since it needs the correlation matrices of all the UEs sharing the same pilot (Björnson; Hoydis; Sanguinetti, 2017). There are simpler methods, such as least squares, that reduce the accuracy of the estimation while minimizing computational complexity. Despite this, more advanced precoders based on MMSE processing will benefit from more precise channel estimates (Björnson; Hoydis; Sanguinetti, 2017). These complex precoders are necessary since simple Maximum Ratio (MR) processing does not necessarily make a cell-free system more efficient in terms of SE than a cellular small cell systems (Björnson; Sanguinetti, 2020a). In this context, this thesis disregards non-MMSE linear channel estimators, even if in some specific situations they may be a good alternative to reduce the system complexity.

The channel estimates obtained through an MMSE estimator are equal to

$$\widehat{\mathbf{h}}_{l,m,k} = \sqrt{\eta_k\tau_{up}}\mathbf{R}_{l,m,k}\boldsymbol{\Psi}_{l,m,t_k}^{-1}\mathbf{y}_{l,m,t_k}^{\text{pilot}}, \tag{2.33}$$

where $\boldsymbol{\Psi}_{l,m,t_k} \in \mathbb{C}^{N\times N}$ is the correlation matrix of the received signal, which can be calculated as (Demir; Björnson; Sanguinetti, 2021)

$$\boldsymbol{\Psi}_{l,m,t_k} = \mathbb{E}\{\mathbf{y}^{\text{pilot}}_{l,m,t_k}(\mathbf{y}^{\text{pilot}}_{l,m,t_k})^{\text{H}}\} = \sum_{k'\in\mathcal{P}_k} \tau_{up}\eta_{k'}\mathbf{R}_{l,m,k'} + \sigma^2_{ul}\mathbf{I}_{N_{l,m}}. \tag{2.34}$$

In addition to uplink pilots, the system may have downlink ones, which will be used for the UEs to decode the downlink signals with instantaneous channels instead of statistics. These pilots are useful in situations with low channel hardening and can be used just for users with a lower level of channel hardening. This approach increases the system SE without requiring many downlink pilots, which would be additional overhead in the coherence block. The estimation procedure for downlink in the UEs is similar to the uplink one in TRPs (Souza et al., 2022).

2.3.4.2 Uplink transmission and CPU/TRP Received Signal

The signal received by TRP m on SB l is formulated as

$$\mathbf{y}^{\text{ul}}_{l,m} = \sum_{k=1}^{K} \mathbf{h}_{l,m,k}\varsigma^{\text{ul}}_k + \mathbf{n}^{\text{ul}}_{l,m}, \tag{2.35}$$

where $\varsigma^{\text{ul}}_k \in \mathbb{C}$ represents the uplink transmitted signal from user k, and $\mathbf{n}^{\text{ul}}_{l,m} \sim \mathcal{CN}(0, I_{N_{l,m}}\, \sigma^2_{\text{ul}})$ denotes the additive noise at the antenna of TRP m on SB l. In the context of the BTRP functional split, the aggregated received signal available at the CPU from TRP m on SB l is given by

$$\begin{aligned} \mathbf{y}^{\text{BTRP,ul}}_{l,m} &= \left[\mathcal{Q}\left(\mathbf{v}^H_{l,m,1}\mathbf{y}^{\text{ul}}_{l,m}\right), \cdots, \mathcal{Q}\left(\mathbf{v}^H_{l,m,k}\mathbf{y}^{\text{ul}}_{l,m}\right), \cdots, \mathcal{Q}\left(\mathbf{v}^H_{l,m,K}\mathbf{y}^{\text{ul}}_{l,m}\right)\right]^T \\ &= \left[y^{\text{BTRP,ul}}_{l,m,1}, \cdots, y^{\text{BTRP,ul}}_{l,m,k}, \cdots, y^{\text{BTRP,ul}}_{l,m,K}\right]^T, \end{aligned} \tag{2.36}$$

where $\mathbf{v}_{l,m,k}$ signifies the linear combining array applied by TRP m on SB l in its antennas to obtain the signal of user k. Then, the combined signal at the CPU corresponding to the symbol transmitted by the UE k is calculated by

$$y^{\text{BTRP,ul}}_{\text{CPU,k}} = \sum_{l=1}^{L}\sum_{m=1}^{M_l} a_{l,m,k}y^{\text{BTRP,ul}}_{l,m}, \tag{2.37}$$

where $a_{l,m,k} \in \mathbb{C}$ represents the weighting coefficients used in signal combining, aimed at optimizing SE. These coefficients are determined by the CPU based on available channel statistics, as the direct channel estimates are not accessible in the BTRP configuration (Björnson; Sanguinetti, 2020a)(Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

For the BCPU functional split, the aggregated channel detailed in subsection 2.3.2 is considered. In this way, the received signal across all TRPs for user k is expressed as

$$\mathbf{y}^{\text{ul}}_{k} = \sum_{k=1}^{K} \mathbf{h}_{k}\varsigma^{\text{ul}}_k + \mathbf{n}_{\text{ul}}, \tag{2.38}$$

where $\mathbf{n}_{\mathrm{ul}} \sim \mathcal{CN}(0, I_{N_{\mathrm{tot}}}\sigma^2_{\mathrm{ul}})$ denotes the receiver noise across all antennas. Then, the combined signal at the CPU corresponding to the symbol transmitted by the UE k is calculated by

$$y^{\mathrm{BCPU,ul}}_{\mathrm{CPU,k}} = \mathbf{v}_k^H \mathcal{Q}\left(\mathbf{y}_k^{\mathrm{ul}}\right), \tag{2.39}$$

where $\mathbf{v}_k = [\mathbf{v}^T_{1,1,k}, \cdots, \mathbf{v}^T_{1,M_1,k}, \mathbf{v}^T_{2,1,k}, \cdots, \mathbf{v}^T_{2,M_2,k}, \cdots, \mathbf{v}^T_{l,m,k}, \cdots, \mathbf{v}^T_{L,1,k}, \cdots, \mathbf{v}^T_{L,M_L,k}]^T$ is the linear combining array applied to all antennas associated with the aggregated channel $\mathbf{h}_k$, defined in subsection 2.3.2, to obtain the signal of user k (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

The quantization process inside (2.36) and (2.39) employs the AQNM framework delineated in subsection 2.3.3.1, introducing a correlation coefficient and quantization noise. Consequently, the signal model for user k at the CPU can be decomposed in desired signal, inter-user interference, receiver noise, and quantization noise, being represented as

$$\begin{aligned} y_k^{\mathrm{ul}} = & \underbrace{\sum_{l=1}^{L}\sum_{m=1}^{M_l} \alpha^{\mathrm{ul}}_{l,m,k}\mathbf{v}^H_{l,m,k}\mathbf{D}_{l,m,k}\mathbf{h}_{l,m,k}\varsigma^{\mathrm{ul}}_k}_{\text{Desired Signal}} \\ & + \underbrace{\sum_{\substack{i=1\\ i\neq k}}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M_l} \alpha^{\mathrm{ul}}_{l,m,k}\mathbf{v}^H_{l,m,k}\mathbf{D}_{l,m,k}\mathbf{h}_{l,m,i}\varsigma^{\mathrm{ul}}_i}_{\text{Inter-User Interference}} \\ & + \underbrace{\sum_{l=1}^{L}\sum_{m=1}^{M_l} \alpha^{\mathrm{ul}}_{l,m,k}\mathbf{v}^H_{l,m,k}\mathbf{n}^{\mathrm{ul}}_{l,m}}_{\text{Receiver Noise}} + \underbrace{q^{\mathrm{ul}}_k}_{\text{Quantization Noise}} \end{aligned}, \tag{2.40}$$

where $\alpha^{\mathrm{ul}}_{l,m,k}$ is the linear correlation coefficient that quantifies the correlation between the original and quantized uplink signals transmitted via the fronthaul. In the case of BCPU implementation, quantization levels do not vary on a per user basis, and consequently $\alpha^{\mathrm{ul}}_{l,m,k}$ is simplified to $\alpha^{\mathrm{ul}}_{l,m}$, in such a way that $\alpha^{\mathrm{ul}}_{l,m,k} = \alpha^{\mathrm{ul}}_{l,m,k'}\ \forall k' \in \{1, \cdots, K\}$. Furthermore, $\mathbf{n}^{\mathrm{ul}}_{l,m} \sim \mathcal{CN}(0, I_{N_{\mathrm{tot}}}\sigma^2_{\mathrm{ul}})$ is the uplink additive receiver noise at the antennas of TRP m at SB l and q^{ul}_k represents the additive quantization noise affecting the uplink signal of user k, which is calculated as

$$q_k^{\mathrm{ul}} = \begin{cases} \sum_{l=1}^{L}\sum_{m=1}^{M_l} \mathbf{v}^H_{l,m,k}\mathbf{q}^{\mathrm{ul}}_{l,m}, & \text{for BCPU} \\ \sum_{l=1}^{L}\sum_{m=1}^{M_l} q^{\mathrm{ul}}_{l,m,k}, & \text{for BTRP} \end{cases}, \tag{2.41}$$

where $\mathbf{q}^{\mathrm{ul}}_{l,m} \sim \mathcal{CN}(0, I_{N_{l,m}}\alpha^{\mathrm{ul}}_{l,m}(1-\alpha^{\mathrm{ul}}_{l,m}))$ represents the additive quantization noise associated with the quantized antenna signals at TRP m on SB l. Moreover, $q^{\mathrm{ul}}_{l,m,k} \sim \mathcal{CN}(0, \alpha^{\mathrm{ul}}_{l,m,k}(1-\alpha^{\mathrm{ul}}_{l,m,k}))$ corresponds to the additive quantization noise for the quantized combined signals of user k (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

2.3.4.3 Downlink transmission and UE received signal

The received data signal in a UE k from the cell-free network is given by

$$y_k^{\mathrm{dl}} = \sum_{l=1}^{L}\sum_{m=1}^{M_l} (\mathbf{h}_{l,m,k})^H \mathbf{x}_{l,m} + n_k^{\mathrm{dl}}, \tag{2.42}$$

where $n_k^{\mathrm{dl}} \sim \mathcal{CN}(0, \sigma_{dl}^2)$ is additive Gaussian noise and $\mathbf{x}_{l,m} \in \mathbb{C}^{N_{l,m}}$ is the transmitted signal of TRP m on SB l to the user k, that is given by

$$\mathbf{x}_{l,m} = \sum_{i=k}^{K} \alpha_{l,m,k}^{\mathrm{dl}} \mathbf{D}_{l,m,k} \mathbf{w}_{l,m,k} \varsigma_k^{\mathrm{dl}} + \mathbf{q}_{l,m}^{\mathrm{dl}}, \tag{2.43}$$

where $\varsigma_k^{\mathrm{dl}} \in \mathbb{C}$ represents the downlink signal intended for UE k within a coherence block sample, $\mathbf{w}_{l,m,k}$ denotes the precoding vector assigned to user k by TRP m in SB l, and $\alpha_{l,m,k}^{\mathrm{dl}}$ is the linear correlation coefficient reflecting the relationship between original and quantized downlink signals transmitted through the fronthaul. In the case of BCPU implementation, quantization levels do not vary on a per user basis, and consequently $\alpha_{l,m,k}^{\mathrm{dl}}$ is simplified to $\alpha_{l,m}^{\mathrm{dl}}$, in such a way that $\alpha_{l,m,k}^{\mathrm{dl}} = \alpha_{l,m,k'}^{\mathrm{dl}}\ \forall k' \in \{1, \cdots, K\}$. Additionally, $\mathbf{q}_{l,m}^{\mathrm{dl}}$ represents the downlink additive fronthaul quantization noise in TRP m on SB l, and both $\alpha_{l,m,k}^{\mathrm{dl}}$ and $\mathbf{q}_{l,m}^{\mathrm{dl}}$ are integral to the AQNM framework discussed in subsection 2.3.3.1 and $\mathbf{q}_{l,m}$ is given by

$$\mathbf{q}_{l,m}^{\mathrm{dl}} \sim \begin{cases} \mathcal{CN}\left(0, \alpha_{l,m}^{\mathrm{dl}}\left(\alpha_{l,m}^{\mathrm{dl}} - 1\right)\sum_{k=1}^{K} \mathbb{E}\left\{\mathbf{w}_{l,m,k}\mathbf{w}_{l,m,k}^{\mathrm{H}}\right\}\right), & \text{for BCPU} \\ \sum_{i=k}^{K} \mathbf{D}_{l,m,k}\mathbf{w}_{l,m,k} q_{l,m,k}^{\mathrm{dl}}, & \text{for BTRP} \end{cases}, \tag{2.44}$$

where $q_{l,m,k}^{\mathrm{ul}} \sim \mathcal{CN}\left(0, \alpha_{l,m}\left(\alpha_{l,m} - 1\right)\right)$ is the quatization noise of the user k in TRP m on SB l (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020).

The received signal at the UE k can be obtained by applying 2.43 in 2.42, which can be decomposed in desired signal, inter-user interference, receiver noise, and quantization noise, being represented as (Demir; Björnson; Sanguinetti, 2021)(Femenias; Riera-Palou, 2020)

$$\begin{aligned} y_k^{\mathrm{dl}} = & \underbrace{\sum_{l=1}^{L}\sum_{m=1}^{M_l} \alpha_{l,m,k}^{\mathrm{dl}} \mathbf{h}_{l,m,k}^{H} \mathbf{D}_{l,m,k} \mathbf{w}_{l,m,k} \varsigma_k^{\mathrm{dl}}}_{\text{Desired Signal}} \\ & + \underbrace{\sum_{\substack{i=1\\ i\neq k}}^{K}\sum_{l=1}^{L}\sum_{m=1}^{M_l} \alpha_{l,m,i}^{\mathrm{ul}} \mathbf{h}_{l,m,k}^{H} \mathbf{D}_{l,m,i} \mathbf{w}_{l,m,i} \varsigma_i^{\mathrm{dl}}}_{\text{Inter-User Interference}}. \\ & + \underbrace{n_k^{\mathrm{dl}}}_{\text{Receiver Noise}} + \underbrace{\sum_{l=1}^{L}\sum_{m=1}^{M_l} \mathbf{h}_{l,m,k}^{H} \mathbf{q}_{l,m}}_{\text{Quantization Noise}}, \end{aligned} \tag{2.45}$$

2.3.4.4 Combiners and precoders

There are multiple possible combining vectors, the most basic one is the MR combining, where

$$\mathbf{v}_k^{\mathrm{MR}} = \widehat{\mathbf{h}}_k, \tag{2.46}$$

which is a scalable and distributed type of combining. However, MR does not provide any strategy to deal with strong interference scenarios since it does not work with the interfering channels. For these cases, more sophisticated techniques are required (Demir; Björnson; Sanguinetti, 2021). The best and most complex linear combiner is based on MMSE, which accounts for all possible interfering and desired channels, allowing maximum interference cancellation. Global channel knowledge will only happen at centralized processing techniques. In this situation, the MMSE combiner is given by

$$\mathbf{v}_k^{\mathrm{MMSE}} = \eta_k \left(\sum_{k'=1}^{K} \eta_{k'} \mathbf{D}_k \left(\widehat{\mathbf{h}}_{k'} \widehat{\mathbf{h}}_{k'}^H + \mathbf{B}_{k'} \right) \mathbf{D}_k + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{N_{tot}} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k, \tag{2.47}$$

where $\mathbf{B}_k$ is the correlation error matrix that can be separated in $\mathbf{B}_k = [(\mathbf{B}_{1,1,k})^T, \cdots, (\mathbf{B}_{1,m,k})^T, \cdots, (\mathbf{B}_{1,M,k})^T, \cdots, (\mathbf{B}_{l,m,k})^T, \cdots, (\mathbf{B}_{L,M,k})^T]^T$. In such a way that the individual elements $\mathbf{B}_{l,m,k}$ can be calculated as

$$\mathbf{B}_{l,m,k} = \mathbf{R}_{l,m,k} - \eta_k \tau_{up} \mathbf{R}_{l,m,k} \mathbf{\Psi}_{l,m,t_k}^{-1} \mathbf{R}_{l,m,k}. \tag{2.48}$$

where $\mathbf{\Psi}_{l,m,t_k}$ is calculated by 2.34. This combining vector considers desired and all interfering channels, optimizing the user's SE. However, it is not scalable since the equation is dependent on K. See Section 2.3.2 for more details. An alternative to solve this issue is the P-MMSE combining, which is given by

$$\mathbf{v}_k^{\mathrm{P-MMSE}} = \eta_k \left(\sum_{k' \in \mathcal{S}_k} p_i \mathbf{D}_k \widehat{\mathbf{h}}_{k'} \widehat{\mathbf{h}}_{k'}^{\mathrm{H}} \mathbf{D}_k + \mathbf{Z}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{I}_{N_{tot}} \right)^{-1} \mathbf{D}_k \widehat{\mathbf{h}}_k \tag{2.49}$$

where $\mathcal{S}_k = \{k' : \mathbf{D}_k \mathbf{D}_{k'} \neq \mathbf{0}_{N_{tot} \times N_{tot}}\}$ is the set of UEs that are partially served by the same TRPs. Moreover, $\mathbf{Z}_{\mathcal{S}_k}$ is given by

$$\mathbf{Z}_{\mathcal{S}_k} = \sum_{k' \in \mathcal{S}_k} p_{k'} \mathbf{D}_k \mathbf{B}_{k'} \mathbf{D}_k. \tag{2.50}$$

While the P-MMSE combiner is scalable, its computational demands can still be high. An alternative, aimed at reducing computational complexity, is the P-RZF combiner, which simplifies the matrix to be inverted by disregarding $\mathbf{Z}_{\mathcal{S}k}$ and reorganizing 2.50. The P-RZF combining vector is expressed as

$$\mathbf{v}_k^{\mathrm{P-RZF}} = \left[\mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} \left(\widehat{\mathbf{H}}_{\mathcal{S}_k}^{\mathrm{H}} \mathbf{D}_k \widehat{\mathbf{H}}_{\mathcal{S}_k} + \sigma_{\mathrm{ul}}^2 \mathbf{P}_{\mathcal{S}_k}^{-1} \right)^{-1} \right]_{:,1} \tag{2.51}$$

where $[\cdot]_{:,1}$ denotes the operation of only keeping the first column of its matrix argument, $\widehat{\mathbf{H}}_{\mathcal{S}_k} \in \mathbb{C}^{M_{tot} \times |\mathcal{S}_k|}$ contains the stacked vectors $\widehat{\mathbf{h}}_i$ with indices $i \in \mathcal{S}_k$, with the first column being $\widehat{\mathbf{h}}_k$, and $\mathbf{P}_{\mathcal{S}_k} \in \mathbb{R}^{|\mathcal{S}_k|\times|\mathcal{S}_k|}$ is a diagonal matrix containing the transmit powers p_i for $i \in \mathcal{S}_k$, listed in the same order as the columns $\widehat{\mathbf{H}}_{\mathcal{S}_k}$. P-RZF combining effectively uses the pseudo-inverse of the estimated partial channel matrix $\widehat{\mathbf{H}}_{\mathcal{S}_k}$, which is regularized by adding a term $\sigma_{\text{ul}}^2\mathbf{P}_{\mathcal{S}_k}^{-1}$ that accounts for the power levels of UE transmission and noise. The regularization procedure mitigates ill-conditioned matrix inversions and noise amplification. The P-RZF combiner will perform well if the channel conditions of the interfering UEs in $\mathcal{S}_k$ are good.

Centralized combining can result in more signaling and overhead data being sent through the fronthaul links. One of the biggest advantages of MR is the fact that signal processing can be made locally. However, MMSE can also be implemented locally, using only local channel estimates. This implementation may be scalable, called L-MMSE, or non-scalable, called LP-MMSE. Both combining vectors can be respectively calculated as

$$\mathbf{v}_{l,m,k}^{\text{L}-\text{MMSE}} = \eta_k \left(\sum_{i=1}^{K} \eta_i \left(\widehat{\mathbf{h}}_{l,m,i}\widehat{\mathbf{h}}_{l,m,i}^{\text{H}} + \mathbf{B}_{il} \right) + \sigma_{\text{ul}}^2 \mathbf{I}_{N_{m,l}} \right)^{-1} \mathbf{D}_{kl}\widehat{\mathbf{h}}_{l,m,k}, \tag{2.52}$$

$$\mathbf{v}_{l,m,k}^{\text{LP}-\text{MMSE}} = \eta_k \left(\sum_{i\in\mathcal{D}_{l,m}} \eta_i \left(\widehat{\mathbf{h}}_{l,m,i}\widehat{\mathbf{h}}_{l,m,i}^{\text{H}} + \mathbf{B}_{il} \right) + \sigma_{\text{ul}}^2 \mathbf{I}_{N} \right)^{-1} \mathbf{D}_{kl}\widehat{\mathbf{h}}_{l,m,k}, \tag{2.53}$$

where $\mathcal{D}_{l,m}$ is the set of UEs served by TRP m of SB l in accordance with the DCC scheme (Demir; Björnson; Sanguinetti, 2021).

Thanks to the duality between the downlink and uplink operations the precoders can be easily obtained from the combining vectors. In this context, the distributed processing precoders MR, L-MMSE, LP-MMSE is given by

$$\mathbf{w}_{l,m,k} = \sqrt{\varrho_{l,m,k}}\,\frac{\mathbf{v}_{l,m,k}}{\sqrt{\mathbb{E}\left\{\mathbf{v}_{l,m,k}^{\text{H}}\mathbf{D}_{l,m,k}\mathbf{v}_{l,m,k}\right\}}}, \tag{2.54}$$

where $\varrho_{l,m,k}$ is the downlink power transmission coefficient to the user k by TRP m on SB l that will depend on the adopted power control strategy (Demir; Björnson; Sanguinetti, 2021).

The calculation of precoding vectors for centralized processing precoders like MMSE, P-MMSE, and P-RZF is derived from the aggregated combiner $\mathbf{v}_k$. Consequently, the precoder for each user k adopts an aggregated form, represented as $\mathbf{w}_k = [(\mathbf{w}_{1,1,k})^T, \cdots, (\mathbf{w}_{1,m,k})^T, \cdots, (\mathbf{w}_{1,M,k})^T, \cdots, (\mathbf{w}_{l,m,k})^T, \cdots, (\mathbf{w}_{L,M,k})^T]^T$, being calculated as

$$\mathbf{w}_k = \sqrt{\varrho_k}\,\frac{\mathbf{v}_k}{\sqrt{\mathbb{E}\left\{\mathbf{v}_k^{\text{H}}\mathbf{D}_k\mathbf{v}_k\right\}}}, \tag{2.55}$$

where ϱ_k is the downlink power transmission coefficient to the user k that will depend on the adopted power control strategy, which can based on equal power, equal power coefficients, proportional to large scale gains, hardening level compensatory and others for both distributed and centralized precoders (Demir; Björnson; Sanguinetti, 2021).

Lastly, each precoder has a different computational complexity associated with it that will impact the system's total computational complexity, detailed in Section 2.3.3.4. Table 3 presents the number of complex multiplications and divisions for the different scalable precoders discussed on this section, with the exception of the MR precoder, that basically was only the complexity of channel estimation, as those are used directly to make the precoder. All equations in the table for all precoders are derived from Demir, Björnson and Sanguinetti (2021) and the subset $\mathcal{Z}_k = \cup_{i\in\mathcal{S}_k}\mathcal{M}_i$ denotes the TRPs serving the UEs that are in $\mathcal{S}_k$, while subset $\mathcal{A}_{l,m} = \cup_{(l',m')\in\mathcal{M}_k}\mathcal{D}_{l',m'}$ represents the UEs with TRPs in common with those served by TRP l, m. Both $\mathcal{Z}_k$ and $\mathcal{A}_{l,m}$ are utilized to calculate common operations performed only once for each UE k or TRP l, m, such as channel estimation.

Table 3 – Number of complex multiplications and divisions required from the network to perform channel estimation and generate the combining vectors for all UEs in each coherence block for different precoding schemes.

Scheme	**Channel estimation**	**Combining vector computation**
P-RZF	$\sum_{l=1}^{L}\sum_{m=1}^{M_l}\left[N_{l,m}\tau_p + (N_{l,m})^2\right]\left\lvert\mathcal{A}_{l,m}\right\rvert$	$\sum_{l=1}^{L}\sum_{m=1}^{M_l}\left[\frac{1}{2}\left(\lvert\mathcal{A}_{l,m}\rvert^2 + \lvert\mathcal{A}_{l,m}\rvert\right)N\right] + \sum_{k=1}^{K}\left[\lvert\mathcal{S}_k\rvert^2 + \left(\sum_{(l,m)\in\mathcal{M}_k} N_{l,m}\right)\lvert\mathcal{S}_k\rvert + \frac{1}{3}\left(\lvert\mathcal{S}_k\rvert^3 - \lvert\mathcal{S}_k\rvert\right) + \lvert\mathcal{S}_k\rvert\right]$
P-MMSE	$\sum_{k=1}^{K}\left(N\tau_p + N^2\right)\lvert\mathcal{Z}_k\rvert$	$\sum_{k=1}^{K}\left\{\frac{1}{2}\left[\left(\sum_{(l,m)\in\mathcal{Z}_k} N_{l,m}\right)^2 + \sum_{(l,m)\in\mathcal{Z}_k} N_{l,m}\right] + \frac{1}{3}\left[\left(\sum_{(l,m)\in\mathcal{M}_k} N_{l,m}\right)^3 - \sum_{(l,m)\in\mathcal{M}_k} N_{l,m}\right] + \left(\sum_{(l,m)\in\mathcal{M}_k} N_{l,m}\right)^2 + \sum_{(l,m)\in\mathcal{M}_k} N_{l,m}\right\}$
LP-MMSE	$\sum_{l=1}^{L}\sum_{m=1}^{M_l}\left[N_{l,m}\tau_p + (N_{l,m})^2\right]\left\lvert\mathcal{D}_{l,m}\right\rvert$	$\sum_{l=1}^{L}\sum_{m=1}^{M_l}\left\{\frac{1}{2}\left[(N_{l,m})^2 + N_{l,m}\right]\lvert\mathcal{D}_{l,m}\rvert + (N_{l,m})^2\lvert\mathcal{D}_{l,m}\rvert + \frac{1}{3}\left[(N_{l,m})^3 - N_{l,m}\right] + N_{l,m}\right\}$

2.3.4.5 User rate and SE

The achievable rate in mMIMO systems is lower bounded by

$$R_k = \phi B \underbrace{\mathbb{E}\left\{\log_2(1 + \mathrm{SINR}_k)\right\}}_{\text{Ergotic SE}}, \tag{2.56}$$

where ϕ is the pre-log factor, B is the system bandwidth in Hz, and SINR_k is the SINR for user k. The SINR encompasses the power levels of both desired and interfering signals. The pre-log factor ϕ differs based on the operation mode, being $\frac{\tau_{dl}}{\tau_c}$ for downlink and $\frac{\tau_{ul}}{\tau_c}$ for uplink, where τ_{ul} and τ_{dl} specify the number of samples within the coherence block reserved for uplink and downlink operations, respectively. It is crucial to ensure that $\tau_{dl} + \tau_{ul} + \tau_{dp} + \tau_{up} \leq \tau_c$ (Björnson; Hoydis; Sanguinetti, 2017)(Demir; Björnson; Sanguinetti, 2021).

The tightest possible lower bound for the achievable rate is presented in 2.56. Nevertheless, this form requires specific equations for different channel estimators and combining schemes. Moreover, it is difficult to obtain a closed-form equation for this bound. Consequently, most of the mMIMO literature resorts to an alternative bound, known as the Use-and-Then-Forget (UatF) bound, expressed as

$$R_k^{\mathrm{UatF}} = \phi B \log_2(1 + \underline{\mathrm{SINR}}_k), \tag{2.57}$$

where $\underline{\mathrm{SINR}}_k$ is referred as the effective SINR for the user k, being given by

$$\underline{\mathrm{SINR}}_k = \frac{|\mathbb{E}\{\mathrm{DS}_k\}|^2}{\mathbb{E}\{|\mathrm{DS}_k - \mathbb{E}\{\mathrm{DS}_k\}|^2\} + \mathbb{E}\{|\mathrm{IS}_k|^2\} + \mathbb{E}\{|\mathrm{QN}_k|^2\} + \sigma^2_{\mathrm{ul|dl}}}, \tag{2.58}$$

where DS_k represents the desired signal, IS_k denotes interference signals, QN_k signifies fronthaul quantization noise, and $\sigma^2_{\mathrm{ul|dl}}$ is the Additive White Gaussian Noise (AWGN) for either uplink or downlink operations. This formulation disregards channel estimates for signal detection and can be used without specific equations for any combiner or channel estimator. Moreover, the accuracy of the UatF bound approaches that of the tighter bound presented in (2.56) in scenarios characterized by higher channel hardening (Björnson; Hoydis; Sanguinetti, 2017)(Demir; Björnson; Sanguinetti, 2021).

2.4 Chapter summary

This chapter introduced a solid theoretical background on the operation and implementation characteristics of cell-free mMIMO networks, detailing differences from traditional cellular networks and presenting a complete system model.

The architecture of cell-free mMIMO was presented, highlighting its unique advantages over traditional cellular networks. It was discussed that cell-free networks offer a more uniform quality of coverage due to the macro-diversity property that arises from the UC communication paradigm. Additionally, it was pointed out that cell-free mMIMO usually has less channel hardening compared to a cellular mMIMO system, which may result in additional UE side processing and overhead communication within the system.

The presented system model can represent cell-free mMIMO with or without segmented fronthaul. It considers uncorrelated or correlated Rayleigh or Rician fading, scalability aspects with a DCC framework, functional splits, fronthaul limitations, and quantization distortion. The model uses MMSE channel estimates from pilot-based transmissions and can consider operation with and without pilot contamination. With all this information, the presented system model represented downlink and uplink transmissions using P-RZF, P-MMSE, and LP-MMSE combiners/precoders to provide SE figures.

The processing tasks for BTRP and BCPU functional splits, which are common in the cell-free mMIMO literature, were distributed between TRPs and CPUs. It was pointed out that BTRP is ideal for distributed processing implementations like LP-MMSE precoding, while BCPU is ideal for centralized processing implementations like P-RZF or P-MMSE. The detailed task distribution, not previously addressed in research unrelated to this thesis, is a contribution of this work.

A model for the number of GOPS to be performed at TRPs and CPUs was presented. It considered the proposed task distribution for BTRP and BCPU splits, as well as, adaptations from a literature model for individual GOPS values of individual tasks.

The fronthaul limitations aspects of the system model were based on a literature AQNM framework, and the quantification of the distortion effects in the quantized signal to obtain quantization noise was discussed. It was shown how the fronthaul distortion can affect BTRP and BCPU splits, and the required fronthaul bitrate for these approaches was also presented.

The foundational elements outlined in this chapter provide essential knowledge and a comprehensive system model for cell-free mMIMO networks. These are used in developing and analyzing the proposed reliability and economic frameworks, which are detailed in later chapters.

3 Theoretical background required for the proposed reliability and economic evaluation frameworks

To grasp the evaluation frameworks discussed in this thesis, it is essential to have a good understanding of some fundamental concepts. More specifically, a reasonable comprehension of graph theory and continuous-time Markov chains is essential for the reliability framework. Additionally, familiarity with basic concepts in techno-economic models and analysis is necessary for the economic framework. This chapter introduces these concepts and specifically for graphs and Markov chains, explain how they can be utilized to model hardware unavailability in networked systems.

3.1 Graph theory

Graph theory is a branch of mathematics that studies networks of connected objects. Its core concept is the graph, which is essentially a set of objects and the connections between them, being used to model relationships and interactions in complex systems. In this context, graph theory is a very useful tool for determining the complex interconnections among various components within a communication network, which is vital to understanding the impacts of a specific component's failure on all its interconnected counterparts (Diestel, 2017).

3.1.1 Graph basics and fundamental concepts

In graph theory, the objects are referred to as nodes and represent discrete entities such as cities on a map, people in a social network, stations in a transportation system, or, like in this thesis, the different components of an interconnected communication network. Each node is a fundamental part of the graph's structure, serving as a "waypoint" or terminal within the network. The relationships or connections between these nodes are depicted as edges, which can be lines connecting pairs of nodes (Diestel, 2017).

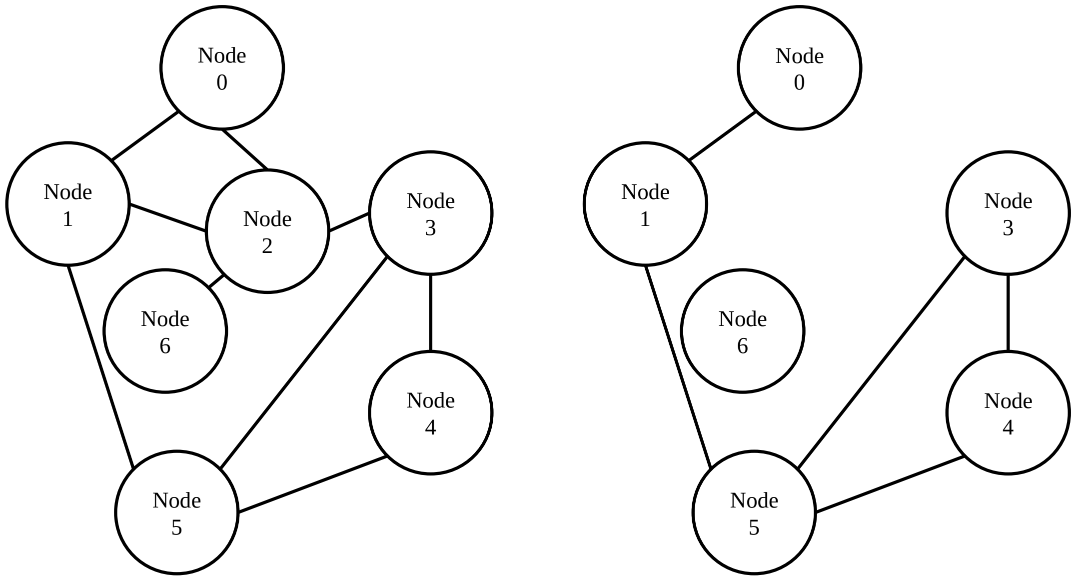
The simplest type of graph is the undirected graph, represented in Figure 15(a). In this graph, there is no inherent direction associated with the connection between any two nodes. Undirected graphs are commonly used to model relationships where mutual interaction is possible, such as in social networks or electrical grids. An easy way to apply an undirected graph to evaluate hardware failure in a communication system is to modify the graph eliminating faulty equipment to look for alternative routes in failure situations (Diestel, 2017).

Suppose a communication network where the nodes are routers that exchange information using unlimited capacity links, considering the shortest number of hops between them. The suppose that Figures 15(a) and 15(b) represent graphs for situations under normal operation and a failure, respectively. Under normal operations, the router at node 3 will connect to node 0 via the router at node 2. In the failed configuration it is visible the absence of node 2 and the four edges associated with it, because this is the failed node. Then, it is easy to see that the path with the fewest hops between nodes 3 and 0 passes through nodes 5 and 1. The same rerouting would happen if the communication link between the routers at nodes 2 and 3 failed and the edge between the nodes was eliminated (Diestel, 2017).

Figure 15 – Illustration of graphs, circles represent nodes, and lines represent edges. The examples contain six to seven nodes and five to nine edges.

(a) Example of undirected graph.

(b) Modification of the example graph removing node 2 and its associated edges.



Source: elaborated by the author.

3.1.1.1 Incidencence, adjacency and reachability

An important graph concept illustrated at the graph in Figure 15(b) is the reachability. Node 6 has no edges and cannot reach any other nodes, and this can happen even if a node has edges. For instance, suppose that the router at node 5 also fails. Then nodes 3 and 4 are reachable among themselves but unreachable from nodes 1, 0, and 6, which are still functioning. In this way we have a disconnected graph with two subgraphs.

Additionally, it is possible to say that a graph is connected when all nodes are reachable between themselves (Diestel, 2017).

Two other important concepts in graphs are: incidence and adjacency. Two nodes are adjacent if there is an edge connecting them, and each edge is incident to the nodes it connects. Thus, two or more nodes may be adjacent to each other, depending on the number of edges in the node, which is called node valency. However, each edge is incident to only the two nodes it connects. The concepts of adjacency and incidence can be respectively represented in matrix form as follows

$$\mathbf{A} = [A_{i,j}] \quad \text{where} \quad A_{ij} = \begin{cases} 1 & \text{if there is an edge between node } i \text{ and node } j \\ 0 & \text{otherwise} \end{cases}, \tag{3.1}$$

$$\mathbf{I} = [i_{i,j}] \quad \text{where} \quad I_{ij} = \begin{cases} 1 & \text{if node } i \text{ is incident to edge } j \\ 0 & \text{otherwise} \end{cases}, \tag{3.2}$$

both equations can be used to evaluate the relationships among nodes and edges for analysis purposes. In an undirected graph, without any self-loop edges at the nodes, the adjacency matrix assumes a symmetrical format concerning the main diagonal (Diestel, 2017). For instance, for the graph of Figure 15(a) the adjacency matrix will be

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.3}$$

3.1.2 Weighted and directional graphs

An effective way to enhance the capabilities of a basic unidirectional graph is to introduce weights on the edges, forming a weighted graph. These weights can represent various factors such as probabilities, transition rates, or distances. Incorporating this additional information enables more sophisticated system modeling and optimization. Weighted graphs have wide-ranging applications across different fields. For example, in transportation network planning, edge weights might signify travel times or costs, aiding in route optimization and infrastructure management. In probabilistic graphical models, weights denote probabilities of transitioning between states, supporting decision-making processes and predictive analytics. (Diestel, 2017).

The previously described rerouting solution, can also incorporate considerations of link capabilities or distances by using a weighted graph. For instance, in a network where routers are represented as nodes, and their connections are depicted in the graph in Figure 15, different rerouting objectives may exist. These objectives could include minimizing the total distance of the new route, maximizing the link capacities to handle increased traffic, or other important operational metrics.

If rerouting aims to minimize distance and node 2 fails, the new connection between nodes 2 and 0 requires summing the edge weights along paths. The fist path include the sequential edges between nodes 3, 4, 5, 1, and 0, and the second path includes the sequential edges between nodes 3, 5, 1, and 0. Moreover, if the weights represent link capacities and the alternate path must have the largest possible capacity, the minimum weight on these sequential paths must be considered. This ensures that the selected path maximizes the minimum weight, providing the greatest capacity for the rerouted traffic.

The main difference in terms of adjacency and incidence matrices for weighted graphs is that they will not be composed of ones and zeros. Instead, these matrices incorporate weights that can denote costs, distances, or probabilities, allowing for more detailed and accurate modeling of various real-world scenarios (Diestel, 2017).

Undirected graphs are valuable but are limited in representing systems where unidirectional processes occur. Directed graphs offer a better solution in such cases, featuring edges with defined directions indicated by arrows pointing from one node to another. This setup is ideal for illustrating asymmetrical relationships, like follower connections on social media or citations between scholarly articles. Each arrow in a directed graph specifies the direction of influence or flow, introducing complexity in pathfinding and other graph-related algorithms. An important difference in terms of adjacency matrix is that symmetry concerning the main diagonal disappears (Diestel, 2017).

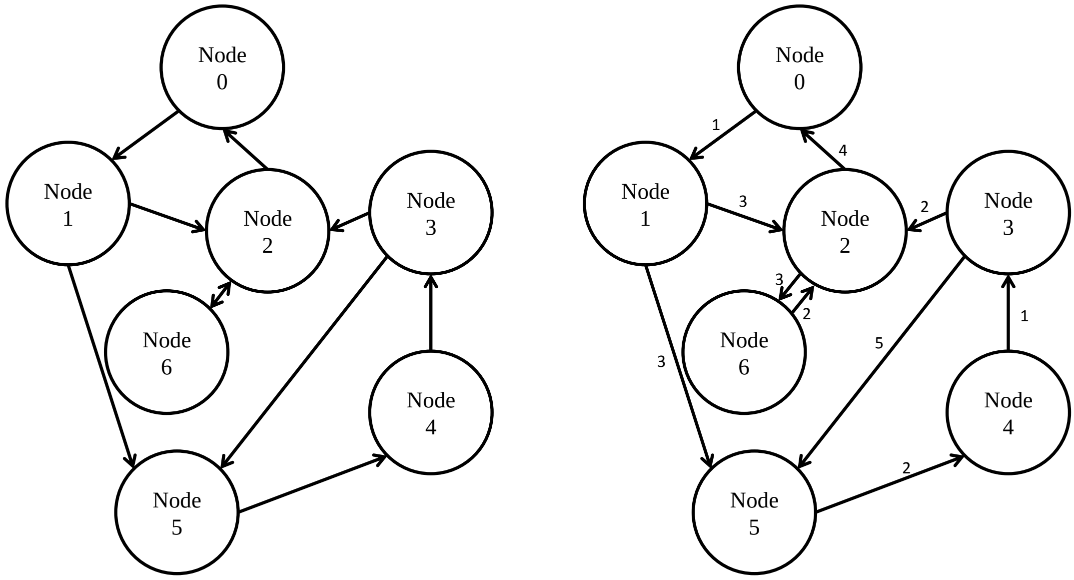
Figure 16 illustrates examples of directed graphs. It is noticeable that in the unweigthed version there is bidirectional edge between Nodes 2 and 6, represented using a line with arrows in both tips. Other possible representation for bidirectionally is to use different two arrows between nodes, each in one direction, as seen in the weighted version. This representation makes more sense in bidirectional graphs with different weights for each direction (Diestel, 2017).

3.1.3 Computational reachability analysis tools for graphs

In the literature, several methods have been developed to verify if nodes are reachable from others. Some of the most common methods are the Depth-First Search (DFS) and the Breadth-First Search (BFS), both of which can explore all the nodes and edges of a graph to determine connectivity. For directed graphs, algorithms such as the Floyd-Warshall algorithm or the transitive closure using matrix multiplication provide systematic ways to determine the reachability of all node pairs. These methods vary in complexity

Figure 16 – Illustrations of a directed graph.

(a) Unweighted. (b) Weighted.



Source: elaborated by the author.

and suitability depending on the size and density of the graph, showcasing the rich toolkit available for handling different graph-related problems in computational contexts (Diestel, 2017)(Goodrich; Tamassia, 2014).

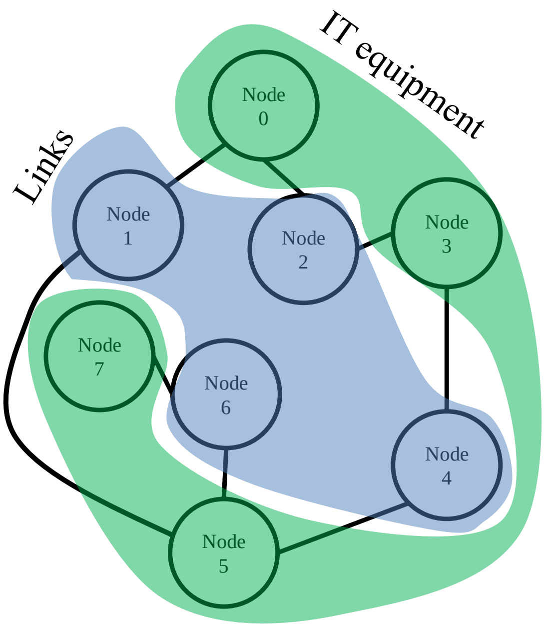
3.1.4 Adaptions for failure representation in networks with redundancy schemes and multi-endpoint links

The examples of graphs used for failure analysis in previous subsections assumed that network Information Technology (IT) components were represented by nodes, and edges represented the connections between them. The issue with this approach is that some protection solutions may create redundant links between nodes or even links connecting more than two IT devices, such as wireless interconnections. Standard graphs typically depict connected nodes with at most one bidirectional edge between them, and an edge can only connect two nodes. Therefore, they may not adequately represent the complexities introduced by protection schemes and multi-endpoint links in communication systems.

One possible solution to this problem is to use hypergraphs, which allow a hyperedge to connect any number of nodes, not just two. However, using hypergraphs makes reachability analysis more complex due to the nature of hyperedges and the richer structure they provide. The multi-node connections in hypergraphs create a more intricate network of relationships, increasing the number of possible paths and making it more difficult to determine reachability. Standard graph algorithms like DFS and BFS require significant adjustments for hypergraphs, necessitating more advanced techniques and computational resources. While hypergraphs offer a powerful way to model complex relationships in networks, their practical application often requires more advanced mathematical and computational tools (Bretto, 2013).

The proposed solution for the problem in this thesis involves using a standard graph where both IT equipment and links are represented as nodes. This approach allows two IT nodes to be connected through multiple link nodes, representing redundancy strategies or the various possible endpoint communications in multi-endpoint links. In this context, edges no longer denote physical transmission mediums but instead symbolize the abstract connections between physical devices and mediums within the communication network. To implement this solution effectively, it is necessary to keep track of the subset of nodes corresponding to IT equipment and those representing links, as shown in Figure 17. Reachability strategies, based on algorithms like DFS and BFS, can be applied to verify which nodes in the IT subset are reachable between themselves.

Figure 17 – Illustration of the considered graph for a failure analysis of communication systems with redundancy schemes and multi-endpoint links.



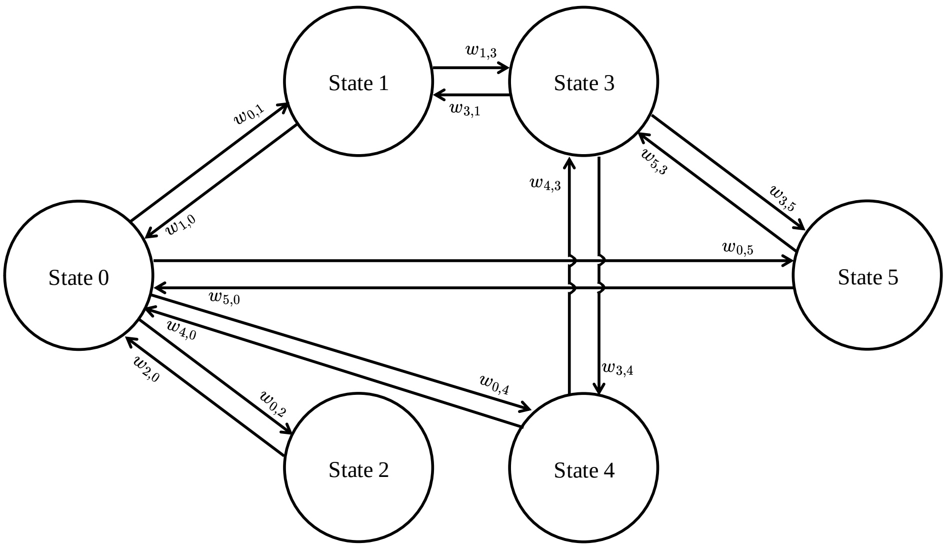
Source: elaborated by the author.

3.2 CTMC for network equipment failure modeling

A Markov process is a type of stochastic process in which the probability of transitioning from the current state to another state depends solely on the present state and not on the sequence of events that preceded it. This characteristic makes them memoryless processes that are very useful in representing a variety of different systems, such as queues, stock market fluctuations, population genetics, physics systems, and even communications network behaviors (Liggett, 2010)(Papoulis; Pillai, 2002).

A Markov chain is a specific type of Markov process with a discrete state space Ω. As a set of predefined states with transition figures between them, the Markov chain can be represented as a weighted graph. Figure 18 presents an example with six states. When the time dependency of being at any state is discrete, the process is referred to as a Discrete-Time Markov Chain (DTMC). When the time dependency of being at any state is continuous, the process is referred to as a CTMC. Markov chains are powerful tools for analyzing and predicting the behavior of complex systems over time (Liggett, 2010)(Papoulis; Pillai, 2002).

Figure 18 – Example of the graph representation of a Markov chain with 6 states, the weights $w_{i,j}$ are transition probabilities or transition rates.



Source: elaborated by the author.

This section focuses on CTMCs as they are an ideal event-driven tool for representing the dynamic behavior of communication networks, where changes can occur at any point in continuous time. CTMCs can effectively model isolated and simultaneous equipment failure states by incorporating appropriate transition rates and incorporating the repair decision into the model. Even the very basic understanding concerning DTMC is provided (Fernandes, 2019)(Fernandez; Stol, 2015)(Fernandez; Stol, 2016).

While the content of this section is not exactly novel to the literature, its presentation aims to contribute by emphasizing reproducibility. Many existing works utilize CTMCs with MCMC solutions to assess failures in communication networks. However, the detailed development of the model is usually abstracted in favor of other analyses (Farias, 2016)(Fernandes et al., 2019)(Fernandes, 2019)(Fernandez; Stol, 2015)(Fernandez; Stol, 2016). This subsection seeks to fill this gap by comprehensively explaining the model development process. Its primary goal is to enhance reproducible research and clarify the application of CTMCs for evaluating failures in communication networks.

3.2.1 DTMC basics

In the case of DTMC, transitions between states occur at fixed discrete time steps. The transition probabilities at the discrete time interval number v are represented in a square matrix with dimensions $|\Omega| \times |\Omega|$ given by

$$\mathbf{P}_v = \begin{bmatrix} p_{1,1,v} & p_{1,2,v} & \cdots & p_{1,j,v} & \cdots & p_{1,|\Omega|,v} \\ p_{2,1,v} & p_{2,2,v} & \cdots & p_{2,j,v} & \cdots & p_{2,|\Omega|,v} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{i,1,v} & p_{i,2,v} & \cdots & p_{i,j,v} & \cdots & p_{i,|\Omega|,v} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ p_{|\Omega|,1,v} & p_{|\Omega|,2,v} & \cdots & p_{|\Omega|,j,v} & \cdots & p_{I,|\Omega|,v} \end{bmatrix}, \tag{3.4}$$

where $p_{i,j,v}$ is the probability of transition from state i to state j at time v. If the transition probabilities do not depend on the step v, the probabilities are said to be stationary, and the chain is homogeneous, simplifying $\mathbf{P}_v$ to $\mathbf{P}$ and $p_{i,j,v}$ to $p_{i,j}$ (Papoulis; Pillai, 2002).

It is important to notice that in homogeneous Markov chains $P(X(n) = i \mid X(m) = j) \neq p_{i,j}$ if $n > m + 1$, i.e., the probability of multiple steps is different from the one-step one. Despite this, the probability of n steps can be calculated from the one-step probability using Chapman-Kolmogorov equation as

$$\mathbf{P}^{(n)} = \mathbf{P}^m \mathbf{P}^{n-m}. \tag{3.5}$$

This equation can be used to perform a large range of analyses with discrete chains (Papoulis; Pillai, 2002).

3.2.2 CTMC basics

In 1931, starting from the theory of discrete-time Markov processes, described by the Chapman-Kolmogorov equation, Andrei Kolmogorov derived a system of first-order differential equations that describe continuous-time Markov processes. Thus, the state probability vector $\mathbf{P}$, with dimensions $|\Omega| \times 1$, for continuous Markov chains can be calculated by:

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{T}, \tag{3.6}$$

where $\mathbf{T}$ is the state transition rate matrix, which can be considered an infinitesimal generator for the stochastic process (Liggett, 2010).

In CTMCs, there is no fixed transition probability matrix, as the transition probability from a state i to a state j changes continuously over time as the system evolves.

Instead, the transitions are governed by the transition rate matrix $\mathbf{T}$, which can be represented as

$$\mathbf{T} = \begin{bmatrix} \lambda_{1,1} & \lambda_{1,2} & \cdots & \lambda_{1,j} & \cdots & \lambda_{1,|\Omega|} \\ \lambda_{2,1} & \lambda_{2,2} & \cdots & \lambda_{2,j} & \cdots & \lambda_{2,|\Omega|} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \lambda_{i,1} & \lambda_{i,2} & \cdots & \lambda_{i,j} & \cdots & \lambda_{i,|\Omega|} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \lambda_{|\Omega|,1} & \lambda_{|\Omega|,2} & \cdots & \lambda_{|\Omega|,j} & \cdots & \lambda_{|\Omega|,|\Omega|} \end{bmatrix} \tag{3.7}$$

where $\lambda_{i,j}$ is the transition rates between states i and j, which may exist for when $i \neq j$. The diagonal elements $\lambda_{i,i}$ of $\mathbf{T}$ are given by

$$\lambda_{i,i} = -\sum_{\substack{j\in\Omega \\ j\neq i}} \lambda_{i,j}. \tag{3.8}$$

The connection between transition probabilities and their rates of change can be described by Kolmogorov's forward and backward differential equations, which are given by

$$\frac{d}{dt}p_{i,j}(t) = \sum_{z} p_{i,z}(t)\lambda_{z,j}, \tag{3.9}$$

$$\frac{d}{dt}p_{i,j}(t) = \sum_{z} \lambda_{i,z}p_{z,j}(t), \tag{3.10}$$

where $p_{i,z}(t)$ and $p_{z,i}(t)$ are state transition probabilities, and the z indexes are the possible intermediate states (Liggett, 2010).

3.2.3 Types of states in a CTMC

In a CTMC, states can be categorized into several types based on their behavior and properties (Anders; Silva, 2000)(Liggett, 2010). Understanding this classification is fundamental for comprehending specific analyses conducted on CTMC. Figure 19 illustrates these state types, and their explanations are as follows

- **Recurrent State**: When starting from this state, the chain will eventually return to it with a probability of 1. Recurrent states exist only in recurrent chains, where all states are recurrent. If the chain is recurrent and all states are reachable from each other, it is known as an irreducible chain. All states in Figure 19(a) exemplify recurrent states.

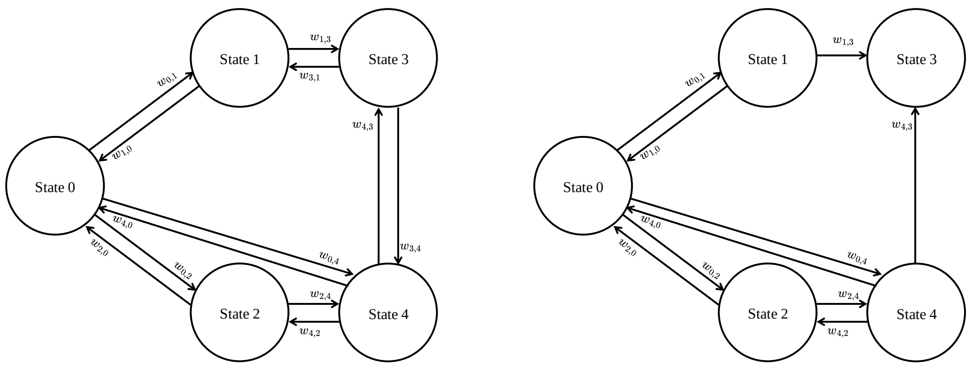
- **Positive Recurrent State**: When starting from this state, the chain will eventually return to it within a finite time frame with a probability of 1. This type of state is a particular case of recurrent states, where the process will return within a reasonable time frame.

- **Absorbing State**: Once the chain enters this state, it cannot leave, remaining there indefinitely. Chains with one or more absorbing states are referred to as absorptive chains. State 3 in Figure 19(b) exemplifies an absorbing state.

- **Transient State**: When starting from this state, there is a non-zero probability that the chain will never return to it. Transient states exist only if at least one absorbing state is in the chain. All states in Figure 19(b), except state 3, exemplify transient states.

Figure 19 – Illustrations of a recurrent and absorptive Markov chain, the example contain 6 states.

(a) Recurrent. (b) Absorptive.



Source: elaborated by the author.

3.2.4 Types of analysis with a CTMC

The analysis of a CTMC can be instantaneous, cumulative, steady-state based, or up-to-absorption. Each type of analysis has unique properties and is useful for different scenarios and applications.

3.2.4.1 Instantaneous analysis

An instantaneous analysis is ideal if it is desired to know the probability of being in any state at any given moment. The probability of being in any state at time t can be calculated by solving the Kolmogorov equations

$$\frac{d\mathbf{P}(t)}{dt} = \mathbf{P}(t)\mathbf{T} \approx \frac{\mathbf{P}(t+\Delta t) - \mathbf{P}(t)}{\Delta t} = \mathbf{P}(t)\mathbf{T}, \tag{3.11}$$

where Δt is an infinitesimal step in time and $\mathbf{P}(t+\Delta t)$ is calculated as

$$\mathbf{P}(t+\Delta t) = \mathbf{P}(t) + \mathbf{P}(t)\mathbf{T}\Delta t = \mathbf{P}(t)(\mathbf{I}_{|\Omega|} + \mathbf{T}\Delta t), \tag{3.12}$$

where $(\mathbf{I}_{|\Omega|} + \mathbf{T}\Delta t)$ is equivalent to the probability transition matrix. From this, it is clear that for a CTMC, the transition rate matrix is the first derivative of the probability transition matrix.

By applying (3.12) in (3.11), the instantaneous state probability vector can be calculated as

$$\mathbf{P}(t) = \mathbf{P}(0)e^{\mathbf{T}t}, \tag{3.13}$$

which involves an exponential matrix function that may be hard to solve analytically, especially for a large space state (Anders; Silva, 2000)(Liggett, 2010).

3.2.4.2 Transient cumulative analysis

A transient cumulative analysis is ideal if one wants to know the time spent in any given state over a period. The array $\mathbf{L}(t) \in \mathbb{R}^{|\Omega|\times 1}$ containing the time spent in each state after a total time t can be calculated by integrating 3.11 over the desired amount of time

$$\mathbf{L}(t) = \int_0^t \mathbf{P}(0)e^{\mathbf{T}\tau}\,d\tau = \frac{\mathbf{P}(0)}{\mathbf{T}}e^{\mathbf{T}t} - \frac{\mathbf{P}(0)}{\mathbf{T}}e^{\mathbf{0}_{|\Omega|}}, \tag{3.14}$$

which has a first derivative equal to

$$\frac{d\mathbf{L}(t)}{dt} = \mathbf{P}(0)e^{\mathbf{T}t}. \tag{3.15}$$

Then, considering (3.15), (3.14) can be modified to

$$\mathbf{L}(t) = \frac{1}{\mathbf{T}}\left[\frac{d}{dt}\mathbf{L}(t) - \mathbf{P}(0)\right] \rightarrow \frac{d\mathbf{L}(t)}{dt} = \mathbf{L}(t)\mathbf{T} + \mathbf{P}(0), \tag{3.16}$$

with the initial condition $\mathbf{L}(0) = \mathbf{0}_{|\Omega|\times 1}$, as there is no time spent in any state at time 0 (Anders; Silva, 2000)(Liggett, 2010).

3.2.4.3 Steady-state analysis

Steady-state analysis is ideal if one wants to know the system's long-term behavior, particularly the probabilities of being in each state after a very long period. The existence of a steady-state distribution requires a recurrent chain where every state is positively recurrent. Under this consideration, steady-state is achieved when

$$\frac{d\mathbf{P}(\infty)}{dt} = 0. \tag{3.17}$$

Applying (3.17) to (3.11), the steady-state probability vector is given by (Anders; Silva, 2000)(Liggett, 2010)

$$\boldsymbol{\pi}\mathbf{T} = \mathbf{0}, \tag{3.18}$$

where $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_i, \cdots \pi_{|\Omega|}]^T$ and

$$\sum_{i\in\Omega} \pi_i = 1. \tag{3.19}$$

3.2.4.4 Up-to-absorption Analysis

An up-to-absorption analysis is ideal if one wants to know how long the system will take to reach a specific state. This type of analysis requires at least one absorbing state in the chain. When the system reaches the absorbing state, no transitions happen anymore. Then, the time to reach the absorbing state is determined by

$$\frac{d\mathbf{L}(t)}{dt} = 0. \tag{3.20}$$

By considering the sets of absorbing states Ω_A and transient states Ω_T ($\Omega = \Omega_A \cup \Omega_T$), applying (3.14) to (3.20), and assuming that initially the system is not in a absorbing state, the mean time spent in transient states before absorption is calculated as

$$\frac{d\mathbf{L}(\infty)}{dt} = 0 \rightarrow 0 = \mathbf{L}(\infty)\mathbf{T} = \mathbf{P}(0) \rightarrow \mathbf{z}\mathbf{T}_{trans} = \mathbf{P}_{trans}(0), \tag{3.21}$$

where $\mathbf{P}_{trans}(0)$ and $\mathbf{T}_{trans}$ are the initial probabilities and transition rates of the sub-chain composed only of the transient states. Moreover, $\mathbf{z} = [z_1, z_2, \ldots, z_i, \ldots, z_{|\Omega_T|}]^T$, in such a way that the Mean Time to Absorption (MTTA) is (Anders; Silva, 2000)(Liggett, 2010)

$$\text{MTTA} = \sum_{i \in \Omega_T} z_i. \tag{3.22}$$

3.2.5 Modeling the chain to represent failures in a communication network

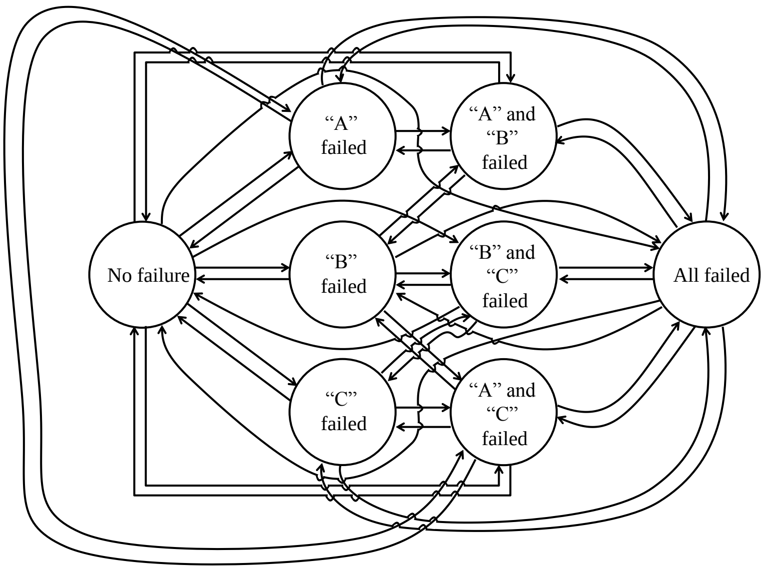
The initial number of possible failed hardware states in a communication network is $2^{Tot_{hard}}$, where Tot_{hard} is the total amount of hardware in the network, accounting for all equipment and individual links. This number represents every possible combination of failures. For instance, a network with 3 hardware pieces will have 1 state with no failures, 3 states with one failure, 3 with a pair of failures, and 1 fully failed scenario. As shown in Figure 20 for the components "A", "B" and "C" (Anders; Silva, 2000)(Fernandes, 2019)(Fernandez; Stol, 2016).

The transition rate between each state depends on the Mean Time Between Failures (MTBF) and Mean Time to Repair (MTTR) of the functional and faulty hardware, respectively. A transition involving one additional failure, equivalent to going from state i to a state j that has only one more piece of equipment failed, can be calculated using the MTBF of all functional equipment in state i. This is given by

$$\lambda_{i,j} = \frac{1}{\text{MTBF}_{(\mathcal{F}_j - \mathcal{F}_i)}} \quad \forall\, |\mathcal{F}_j - \mathcal{F}_i| = 1, \tag{3.23}$$

where $\mathcal{F}_j$ and $\mathcal{F}_i$ are the sets of failed equipment in states j and i, respectively. Transition rates back from state j to state i, where j has one additional failure, depend on the availability of repair teams. The number of possible return transition rates from state j

Figure 20 – Graph representation of an example of a CTMC modeling failed equipment for a network with three distinct components (A, B, and C).



Source: elaborated by the author.

to a state i with one less failure is the minimum between the number of failed equipment in $\mathcal{F}_j$ and the number of available repair teams. The calculation of these rates takes into consideration the MTTR of failed devices and is given by

$$\lambda_{j,i} = \frac{1}{\mathrm{MTTR}_{(\mathcal{F}_j - \mathcal{F}_i)}} \quad \forall\, |\mathcal{F}_j - \mathcal{F}_i| = 1. \tag{3.24}$$

Transition rates between states with more than ones pieces of equipment failing, i.e., $|\mathcal{F}_j - \mathcal{F}_i| > 1$, are more complex to obtain. They involve determining the probability of more than one equipment failing simultaneously over a period of time $(p^{0\to t}_{\mathcal{F}_j - \mathcal{F}_i})$[1]. It is known that the probability of failure relates to reliability as its complementary probability, in this way

$$R^{0\to t}_{\mathcal{F}_j - \mathcal{F}_i} = 1 - p^{0\to t}_{\mathcal{F}_j - \mathcal{F}_i}. \tag{3.25}$$

Reliability (R_s) is related to failure rate (λ) by

$$R_\mathrm{s} = e^{-\lambda t}, \tag{3.26}$$

$$\lambda = -\frac{\ln(R_\mathrm{s})}{t}, \tag{3.27}$$

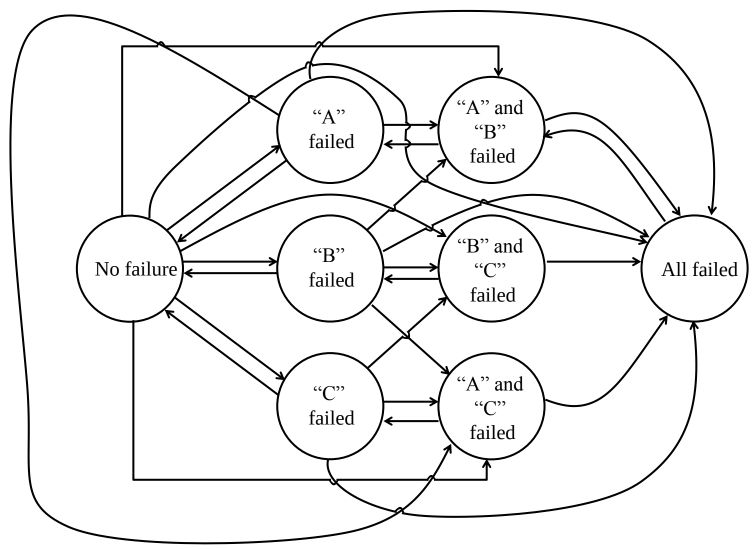
which can be used with Equation (3.25) to obtain failure rates when more than one additional failure occurs (Kletz, 2001).

[1]While the probability of two pieces of equipment failing at the exact same time can be very small, excluding environmental impacts and cascading destruction effects, there are instances where failures occurring at different times can effectively be treated as simultaneous. For example, this can happen when the minimum time required to initiate a response for repair overlaps with the occurrence of another potential failure.

For backward repair transitions concerning more than one additional failure, the same equations apply but with the probability of more than one equipment being repaired simultaneously over a period of time. The main difference is that the number of total failures that can be repaired simultaneously is proportional to the number of repair teams, allowing the system to return to states where $|\mathcal{F}_j - \mathcal{F}_i|$ is equal to or smaller than the number of repair teams. The choice of which equipment will be repaired at each state can vary according to different parameters, but usually the first repair team prioritizes equipment whose repair will most significantly mitigate the failure's impact proportionally to the repair time. Subsequent teams will similarly prioritize the next most critical equipment for repair.

Figure 21 modifies the previously presented chain in Figure 20 to account for the existence of just one repair team. It is noticeable that all states with multiple failures have just one possibility of repair, predefined on the chain, because it was considered the more advantageous repair. In this sense, Markov chains incorporate the repair decision process into their model (Fernandez; Stol, 2015)(Fernandez; Stol, 2016).

Figure 21 – Graph representation of an example of a CTMC modeling failed equipment for a network with three distinct components (A, B, and C) and only one repair team.



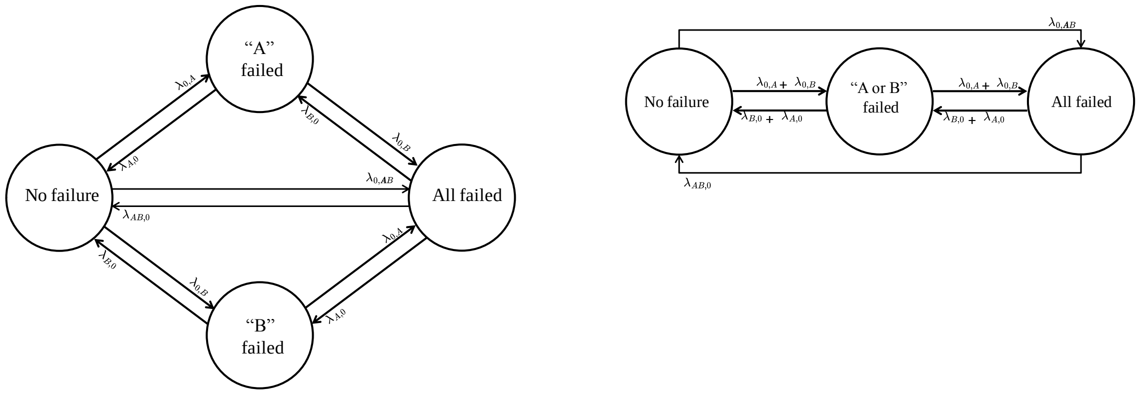
Source: elaborated by the author.

3.2.5.1 Chain reduction

Sometimes, it is possible to reduce the Markov chain depending on the desired type of analysis. For instance, consider a network with two identical pieces of equipment, "A" and "B". This network will have four states, as represented in Figure 22(a). However, a failure in "A" will have the same impact as a failure in "B", allowing the states equivalent to the failures of those devices to be considered identical. Consequently, the chain can be simplified to the one shown in Figure 22(b), where the failure rate of the new state is simply the sum of the failure rates of the previous states. Notably, this latter chain is irreducible, as all states can connect to each other (Anders; Silva, 2000)(Liggett, 2010).

Figure 22 – Example of an reduction in a CTMC modeling failed equipment for a network with two equal components (A and B).

(a) Original chain. (b) Reduced chain.



Source: elaborated by the author.

It is not always possible to make the association mentioned, even with the same type of equipment. For example, identical devices installed in different locations may have different repair travel times, preventing complete similarity. Despite this, it is sometimes possible to make such associations due to symmetries in specific scenarios.

Whenever a new state is formed as an "or" operation among different states, the failure rate can be represented by a summation, the failure rate of the state j representing the failure of one additional hardware concerning a state i is given by

$$\lambda_{i,j} = \sum_{h \in \mathcal{H}_{i \to j}} \frac{1}{\mathrm{MTBF}_h} \tag{3.28}$$

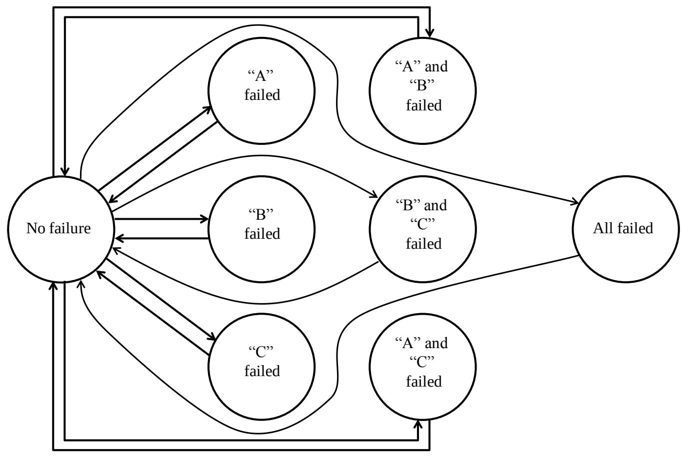
where MTBF_h is the MTBF of device h and the set $\mathcal{H}_{i \to j}$ represents the possible hardware that can fail at a state i, in such a way that $\mathcal{H}_{i \to j} \cap \mathcal{F}_i = \emptyset$.

Equal equipment and characteristics are not the only reasons to aggregate states. Often, the aggregation depends on the type of desired analysis. Figure 23(a) presents a CTMC where a state with no failures can transition to any possible failure state, such that no other transitions occur except for repairs back to the no-failure state. This chain is ideal for verifying the probability of transitioning from a no-failure configuration to a state with 1 to Tot_{hard} failed components. To this end, the chain can be simplified to the format in Figure 23(b).

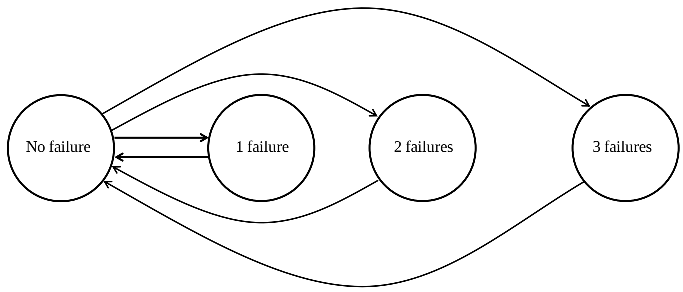
While "or" aggregation of states is easily performed by summing transition rates, there is no straightforward way to perform state association under an "and" operation. Therefore, this type of operation is not used to reduce chains (Anders; Silva, 2000)(Kletz, 2001)(Liggett, 2010).

Figure 23 – Example of an reduction in a CTMC modeling transitions between a no-failure state to failed configurations. Three different equipment are considered.

(a) Original chain.



(b) Reduced chain.



Source: elaborated by the author.

3.2.5.2 Approximations for a network with very reliable devices

Many communication networks feature highly reliable individual equipment, with very low probabilities of failures. MTBFs higher than 50,000 hours are almost an norm, and values around 300,000 to 1,000,000 hours are very common.

Therefore, the probability of failure of a single device is very low, resulting in low probabilities for simultaneous failures happening at the exact same time if the failures of two devices are independent events. This fact suggests that transition rates between states i and j with more than one different equipment failed, i.e., $|\mathcal{F}_j - \mathcal{F}_i| > 1$, can be disregarded. This does not mean that a two-equipment failed state cannot be achieved from the no-failure state, just that it is very improbable to achieve this state directly from the no-failure state.

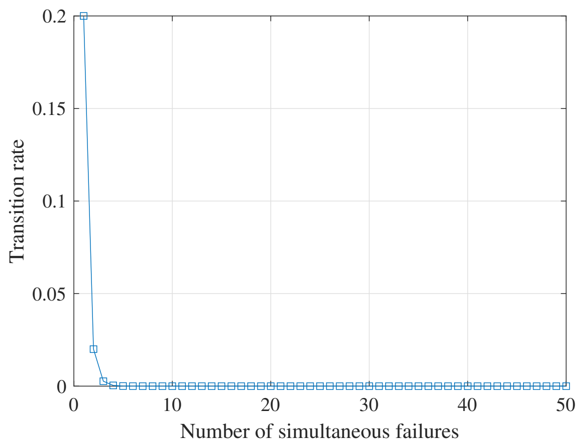
One might argue against this simplification by noting that the number of possible states increases significantly when considering simultaneous failures under higher equipment counts. For instance a network with 10 equipments will have 120 states related to 3 failures occurring concomitantly, and 252 states for 5 failures occurring concomitantly. Then, significantly higher failure rates can be obtained for simultaneous failures compared to individual failures when making the reduction in Figure 23. Therefore, one can conclude that proportion of failure rates related to simultaneous failures cannot be disregarded, rendering the aforementioned simplification invalid.

Despite this, the reliability of individual equipment in communication networks is typically so high that even with a large number of states of large networks, simultaneous failures transitions can often be disregarded.

For instance, consider a network with 10,000 pieces of equipment, each with a 50,000-hour MTBF (relatively low for communication network equipment), where failures are deemed simultaneous if they occur within a 1-hour interval (a generous timeframe). In this situation, the reliability of individual equipment can be calculated using 3.26 with $\lambda = \frac{1}{50{,}000}$ and $t = 1$. Then, the reliability when multiple devices fail within this timeframe is given by $1 - (1 - R_s)^{|\mathcal{F}_j - \mathcal{F}_i|}$, and the resultant associated failure rate for simultaneous failures is obtained using 3.27. Finally, the failure rates of states with the same count of failed equipment are summed to form a chain similar to the one in Figure 23(b).

Under these considerations, Figure 24 illustrates the aggregated failure rate for up to 50 simultaneous failures. It is evident that the sum of the failure rates for individual failures is significantly higher compared to simultaneous failures, accounting for approximately 90% of the total summed failure rates. Increasing the MTBF to 300,000 hours (an very common value in many times of communication equipment) increases this figure to approximately 99%.

Figure 24 – Aggregated failure rate for up to 50 simultaneous failures for a system with 10000 equipments, each with a 50000 hours MTBF, and chain similar to the one in Figure 23.



Source: elaborated by the author.

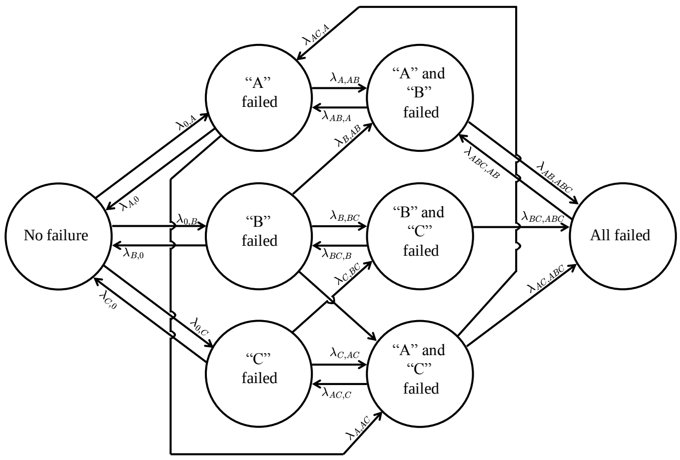
While simultaneous failure transition rates may be disregarded, the repair ones cannot, given that MTTR is typically much smaller than MTBF. The exception is when the system has just one repair team. Outside of this situation and considering the set of equipment to be repaired from state j to i as $\mathcal{R}_{j,i}$, the transition rate for repairing multiple devices simultaneously can be calculated as

$$\lambda_{j,i} = -\frac{\ln\left[1 - \prod_{z \in \mathcal{R}_{j,i}} \left(1 - e^{-\frac{1}{\mathrm{MTTR}z}\mathrm{MTTR}_{\min}}\right)\right]}{\mathrm{MTTR}_{\min}}, \tag{3.29}$$

where $\mathrm{MTTR}_{\min}$ is the minimum MTTR among the set of equipment to be repaired. This equation is derived from Equations (3.26) and (3.27), utilizing $\mathrm{MTTR}_{\min}$ and the relevant timeframe.

Based on the content of this subsection the chain at Figure 21 can be simplified to the one in Figure 25, by assuming only 1 step failure transitions and one repair team, it is notable that the new chain is way less complex than the original one in Figure 20.

Figure 25 – Graph representation of an example of a CTMC modeling failed equipment for a network with three distinct components (A, B, and C) under a 1 step failure simplification and only one repair team.



Source: elaborated by the author.

3.2.6 Monte Carlo simulation approaches for cumulative and absorption analysis

Even the instantaneous solution of a CTMC involves an exponential matrix function that may be hard to calculate, especially when the number of states is very high. In this way, numerical and simulation approaches are common to perform analysis with CTMC. The most useful analyses are the cumulative and up-to-absorption ones. The first can determine the system behavior for any period of time, be it short or long, which will tend to a steady-state analysis. The latter is very useful to calculate the time until a certain state or condition is achieved on the network, which can be a very important measure for network design.

Both these analyses can be performed through Monte Carlo simulation, which is very flexible and allows any temporal aspects of system operation to be easily incorporated into the model. An essential aspect for both analyses is the expected holding time for the CTMC in any state i, which is given by

$$T_i = \frac{1}{|\lambda_{i,i}|}, \tag{3.30}$$

where the $|\cdot|$ operation is used to compensate for the negative value of $\lambda_{i,i}$. Another essential aspect for a Monte Carlo simulation is the probability of transition to a state j from a state i, which is given by

$$P_{i\to j} = \begin{cases} \frac{\lambda_{i,j}}{|\lambda_{i,i}|}, & \text{for } i \neq j \\ 0, & \text{otherwise} \end{cases}, \tag{3.31}$$

which can be organized in the array $\mathbf{P}_i^{\text{from}} = [P_{i\to1}, P_{i\to2}, \cdots, P_{i\to j}, \cdots, P_{i\to|\Omega|}]$, in such a way that $\sum_{j\in\Omega} P_{i\to j} = 1$ (Anders; Silva, 2000).

Then, the cumulative analysis and up-to-absorption analysis are performed according to the algorithms 1 and 2, where N_{trials} is the number of trials of the Monte Carlo simulation. Ideally, this number should be pretty high to ensure a convergent behavior. The algorithms 1 and 2 assume a prefixed value, but implementations can be made based on the variation of the desired output over the iterations. in such a way that when the variation is bellow a minimum threshold no trials should be executed anymore (Anders; Silva, 2000)(Fernandes, 2019).

3.3 Cost consideration with Markov reward models

Markov reward models extend CTMC analyses by associating rewards with states or transitions. This allows for more analyses than just probabilities, time spent in states, and MTTA. One of the aspects that can be considered under reward models are cost considerations (Fernandez; Stol, 2015).

Rewards for transitions involve assigning a reward to each shift from a state i to state j. A reward matrix $\mathbf{R} = [R_{i,j}] \forall i \in \Omega$ and $j \in \Omega$, is obtained for each transition rate $\lambda_{i,j}$ in$\mathbf{T}$. An important metric in this context is the frequency of transitions between two states, which is calculated from the frequency of accesses to a state i, given by

$$f_i = \frac{\pi_i}{T_i}, \tag{3.32}$$

which can be used with $\lambda_{i,j}$ to calculate the expected frequency of the transition from state i to state j

$$F_{i,j} = P_{i\to j} f_i = \lambda_{i,j}\pi_i. \tag{3.33}$$

In cumulative analysis with computational solutions like the ones in algorithm 1, it is necessary to have a matrix to keep track of the number of accesses from each state i

Algorithm 1: CTMC cumulative analysis using Monte Carlo simulations.

Input: $\mathbf{P}(0)$, $\mathbf{T}$, N_{trials},$|\Omega|$
T_STATE $\leftarrow [0]_{1\times|\Omega|}$ ▷ Initializes an array of zeros representing the spent at each state
for *trial* $\leftarrow 1$ **to** N_{trials} **do**
  CURRENT_STATE ← OBTAIN_INITIAL_STATE($\mathbf{P}(0)$) ▷ Calculates initial state using $\mathbf{P}(0)$ as a discrete probability density function
  CURRENT_TIME ← $T_{\text{CURRENT\_STATE}} \ln(U)$ ▷ Based on equation (3.30), U is a pseudo-random variable with a uniform distribution
  T_STATE[*Current_state*] ← *Current_time*
  while *CURRENT_TIME* < *TARGET_TIME* **do**
    NEXT_STATE ← OBTAIN_NEXT_STATE(CURRENT_STATE, $\mathbf{T}$) ▷ Based on equation (3.31) $\mathbf{P}_i^{\text{from}}$ is obtained from $\mathbf{T}$ and CURRENT_STATE, being used as a discrete probability density function to obtain the next state
    NEXT_TIME ← $T_{\text{NEXT\_STATE}} \ln(U)$ ▷ Based on equation (3.30), U is a pseudo-random variable with a uniform distribution
    T_STATE[NEXT_STATE] ← **T_STATE**[NEXT_STATE] + NEXT_TIME
    CURRENT_STATE ← NEXT_STATE
    CURRENT_TIME ← CURRENT_TIME + NEXT_TIME
  end
end
T_STATE ← **T_STATE**/N_{trials}
Output: T_STATE

to each state j. This matrix is updated during the trials and normalized by the number of trials at the end. Under these considerations, the costs of the system can be calculated by a Markov reward model equivalent to $\sum_{i\in\Omega}\sum_{j\in\Omega} C_{i,j}^{MkRw}$ where

$$C_{i,j}^{MkRw} = A_{i,j}^{\text{Total}} r_{i,j}, \tag{3.34}$$

where $A_{i,j}^{\text{Total}}$ is the total expected number of accesses from state i to j during the analysis period.

By knowing the number of accesses or the frequency of access, the total reward for the system is calculated by multiplying the frequency by the reward values. Rewards to transitions are less common than rewards to states but can be used in specific situations. For example, suppose that a company pays a bonus to its technicians for different types of repairs. In this situation, the bonus cost is not dependent on the final replied state but on what kind of repair is made, which in a chain like the one in Figure 25 would be the transition.

Rewards for transitions are the usual type of reward adopted, they can consider steady state probabilities or cumulative time. Essentially each state has a reward to it and the all rewards can be represented in the array $\mathbf{r} = [r_1, \cdots, r_i, \cdots, r_{|\Omega|}]^T$. In terms of cost, usually operating at different states will result in different costs, associated with different energy consumption or penalty costs. Considering the cumulative analysis, by knowing

Algorithm 2: CTMC up-to-absorption analysis using Monte Carlo simulations.

Input: $\mathbf{P}(0)$, $\mathbf{T}$, N_{trials},$|\Omega|$,Ω_A

T_STATE $\leftarrow [0]_{1\times|\Omega|}$ ▷ Initializes an array of zeros representing the spent at each state
for *trial* $\leftarrow 1$ **to** N_{trials} **do**
    CURRENT_STATE $\leftarrow$ OBTAIN_INITIAL_STATE($\mathbf{P}(0)$) ▷ Calculates initial state using $\mathbf{P}(0)$ as a discrete probability density function
    CURRENT_TIME $\leftarrow T_{\text{CURRENT\_STATE}} \ln(U)$ ▷ Based on equation 3.30 ,U is a pseudo-random variable with a uniform distribution
    T_STATE[*Current_state*] $\leftarrow$ *Current_time*
    while *CURRENT_STATE* $\notin \Omega_A$ **do**
        NEXT_STATE $\leftarrow$ OBTAIN_NEXT_STATE(CURRENT_STATE, $\mathbf{T}$) ▷ Based on equation 3.31 $\mathbf{P}_i^{\text{from}}$ is obtained from $\mathbf{T}$ and CURRENT_STATE, being used as a discrete probability density function to obtain the next state
        if *NEXT_STATE* $\notin \Omega_A$ **then**
            NEXT_TIME $\leftarrow T_{\text{NEXT\_STATE}} \ln(U)$ ▷ Based on equation 3.30 ,U is a pseudo-random variable with a uniform distribution
        end
        T_STATE[NEXT_STATE] $\leftarrow$ **T_STATE**[NEXT_STATE] + NEXT_TIME
        CURRENT_STATE $\leftarrow$ NEXT_STATE
    end
end
T_STATE $\leftarrow$ **T_STATE**/N_{trials}
MTTA $\leftarrow$ SUM_ELEMENTS(**T_STATE**)

Output: MTTA

the total time spend at each state over a period t. Under these considerations, the costs of the system can be calculated by a Markov reward model equivalent to $\sum_{i\in\Omega} C_i^{MkRw}$ where

$$C_i^{MkRw} = T_i^{\text{Total}} r_i, \tag{3.35}$$

where T_i^{Total} is the total expected spent time in state i during the analysis period (Fernandez; Stol, 2015).

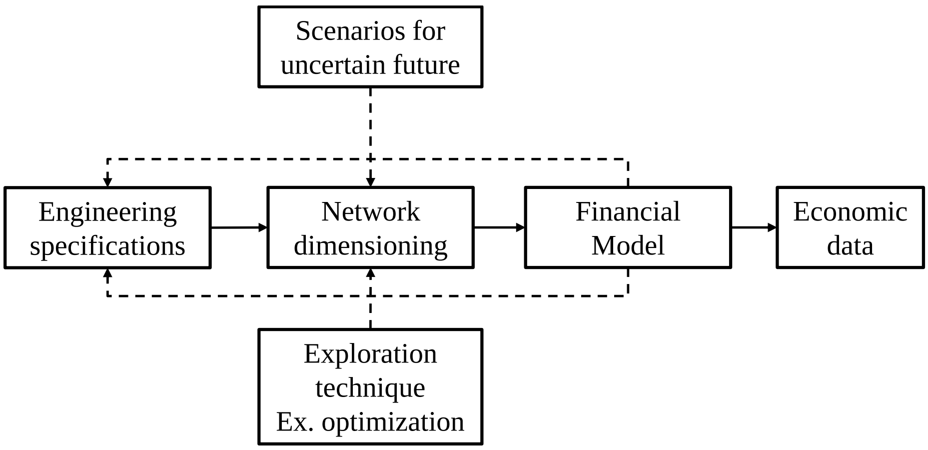
3.4 Techno-economics assessments

Techno-economics is used in engineering to evaluate the economic performance and impacts of engineered systems and projects. This evaluation can use quantitative and qualitative modeling techniques. This section, and subsequently the thesis, specifically focuses on the quantitative approach, which utilizes input values and equations for assessing the financial viability, cost-effectiveness, and potential economic returns of different engineered systems. Additionally, these analyses can measure other important impacts on the systems and projects, such as energy efficiency, carbon footprint, and overall sustainability.

3.4.1 Techno-economics for communication networks

In the context of communication networks, quantitative techno-economics have been studied of a wide range of technologies, including Third Generation of Mobile Systems (3G), 4G, 5G, Wireless Fidelity (Wi-Fi), satellite broadband, Heterogenous Networks (HetNets), backhaul networks, industrial networks, wireless spectrum sharing, fixed access broadband and others. Figure 26 provides a general theoretical overview of a techno-economic analysis applied to a wireless communication network. The model integrates engineering specifications, dimensioning representation, and financial aspects to obtain economic data. Additionally, the model can accommodate variations in scenarios to address uncertain futures and employ model techniques such as optimization (Farias, 2016)(Oughton; Lehr, 2022)(Yaghoubi et al., 2018).

Figure 26 – Overview of a techno-economic assessment for a wireless communication network.

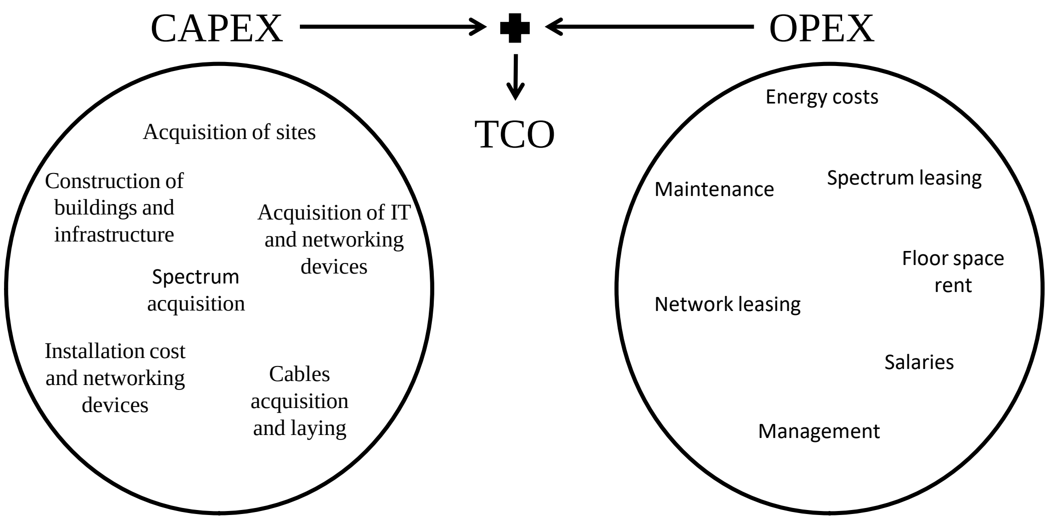


Source: adapted from (Oughton; Lehr, 2022).

The lifecycle of a communication network typically consists of three main phases: planning, installation, and operation. An additional phase, migration, may be needed when upgrading a network in the middle of its lifecycle. To the operator, the planning phase is crucial for ensuring profits and reducing investment risk. This phase requires quantifying the total expenses needed during the network's lifecycle and estimating the Total Cost of Ownership (TCO) over its expected operational time. TCO combines both CAPEX and OPEX to provide a complete understanding of the engineered systems' financial aspects, being crucial for comparing different technological options and choosing the best balance between performance and cost over time (Yaghoubi et al., 2018).

CAPEX and OPEX encompass different costs involved in system deployment and operation, as illustrated in Figure 27. CAPEX refers to upfront expenses incurred for acquiring, upgrading, and deploying physical assets. In communication networks, this includes expenditures on infrastructure such as BSs, antennas, servers, and other hardware essential for establishing the network. CAPEX represents a significant, one-time investment concerning deployed infrastructure that establishes the network's foundation. It is important to note that if migration is adopted, CAPEX can occur at any point during the network's lifecycle. This fact may seem unusual as it suggests CAPEX is not strictly a one-time investment. However, the crucial point is that the investment for the additional infrastructure for the migration is a one-time event. When migration phases are disregarded, i.e., the network is planned to support all possible demands in its lifecycle, CAPEX typically happens within the first year of operation. On the other hand, OPEX encompasses ongoing costs necessary to operate and maintain the engineered system. For communication networks, OPEX includes expenses such as electricity, maintenance, labor, network site leasing, and network management and administration costs. Unlike CAPEX, OPEX is recurring and must be effectively managed to ensure the network's long-term sustainability and profitability (Farias, 2016)(Souza et al., 2021)(Yaghoubi et al., 2018).

Figure 27 – Examples of costs related to CAPEX and OPEX in wireless networks.



Source: elaborated by the author.

There is no set method for developing cost equations for CAPEX and OPEX components. The only requirement is that the equations should accurately represent the cost components, confirmed by market research. Different types of equations are often utilized within the same techno-economic model to compute various costs. Some equations are linear, containing just simple summations and products. The global power consumption may be calculated in this way by summing the products of the power consumption of the devices by the price per Wh. Other equations involve non-linear elements, such as pricing models for devices and infrastructure. These models might include minimum step prices related to scaling requirements, different operational options indicated by binary variables, and non-linear cost curves based on market research, valid across different price ranges and requirements. Complex non-linear equations may improve precision but may have a marginal impact on overall costs depending on the analyses, while simple formulas can be effective without compromising the integrity of the model. Despite this, certain analyses may require more complex equations or simulations. For instance, when evaluating failure penalty costs in communication systems with production redundancy schemes, Monte Carlo simulations are often utilized due to the numerous possible states of connections and failures (Farias, 2016)(Fernandez; Stol, 2015)(Fernandez; Stol, 2016)(Oughton; Lehr, 2022).

3.4.2 Additional relevant financial metrics

While TCO is an essential metric in evaluating overall costs and feasibility, the profitability of any investment is evaluated by different metrics like the Return on Investment (RoI) or the Net Present Value (NPV). The first is a straightforward metric that measures the efficiency of an investment by comparing the net profit to the initial investment cost. Then, the RoI of a communication system with K users after a large CAPEX investment can be expressed in percentage as

$$\mathrm{RoI} = \frac{K\lambda_{\mathrm{sub}}T_{\mathrm{ope}}^{\mathrm{sub}} - \mathrm{OPEX}}{\mathrm{CAPEX}} \times 100, \tag{3.36}$$

where λ_{sub} is the subscription price and $T_{\mathrm{ope}}^{\mathrm{sub}}$ is the total operational time in the time unit of the subscription. While RoI is helpful in checking if an investment yields returns relative to the initial capital, it fails to consider cash discount rates. Those reflect the principle that money today is worth more than the same amount in the future due to its potential earning capacity or inflation effects. Then, to adequately analyze the profitability of any engineered system, it is better to consider its NPV, which is given as

$$\mathrm{NPV} = \sum_{i=0}^{N_{\mathrm{ti}}} \frac{C_i}{(1+r)^i}, \tag{3.37}$$

where C_i represents the cash flow at time i and time is discretely in N_{ti} intervals, such as in years. The value of C_i can be calculated as

$$C_i = K\lambda_{sub}T_i^{\mathrm{ope}} - \mathrm{TCO}_i, \tag{3.38}$$

T_i^{ope} is the time spent at the discrete time interval i and TCO_i is the TCO at this interval. TCO_0 is essentially the base CAPEX cost to deploy the network. A positive NPV indicates that the projected earnings exceed the anticipated costs and that the deployment is profitable (Pärssinen et al., 2019)(Yaghoubi et al., 2018).

Another common metric for economic evaluation is the Internal Rate of Return (IRR), which is the discount rate that makes the NPV of all cash flows from a particular project equal to zero for a given time period. The IRR is found by solving as

$$0 = \sum_{i=0}^{N_{\mathrm{ti}}} \frac{C_i}{(1+\mathrm{IRR})^i}, \tag{3.39}$$

which is usually solved in a numeric way (Pärssinen et al., 2019).

In this thesis, it is also considered a Breakeven Subscription Value (BSV) metric, which is inspired by the IRR. The main difference is that the subscription value that ensures the NPV of all cash flows from the deployment is equal to zero, being calculated by

$$0 = \sum_{t=0}^{N_{\mathrm{ti}}} \frac{K \times \mathrm{BSV} \times T_i^{\mathrm{ope}} - \mathrm{TCO}_i}{(1+r)^i}, \tag{3.40}$$

which is usually solved in a numerical way.

In summary, the complete TCO is essential to identify any cost behavior in an engineered system, considering only CAPEX or OPEX would hide important cost behaviors for the planning phase of any communication network. Then, by using the TCO, the RoI, NPV, IRR and BSV can be evaluated in such a way that:

- If RoI is higher than zero, the required rate of return, the investment is considered efficient. RoI provides a relative measure in percentage terms.
- If NPV is higher than zero, the investment is considered profitable. It gives an absolute measure of the expected dollar value increase.
- If IRR is higher than the required rate of return, the investment is considered good, mainly because it will be profitable even with discounts higher than the ones required.
- If BSV is smaller or equal to the maximum acceptable subscription fee, the investment is considered good.

3.5 Chapter summary

This chapter presented the basic concepts required to understand the proposed reliability and economic evaluation frameworks, which are detailed in later chapters. It explains how graphs and Markov chains can be utilized to model hardware unavailability in networked systems, as well as basic cost considerations in a communications network.

For graph theory, the ideas of undirected, directed, and weighted graphs, as well as the concepts of incidence, adjacency, and reachability, were provided, along with examples for failure modeling in communication networks. These concepts are essential even when representing networks with redundancy schemes or multi-endpoint links.

Regarding CTMCs, the types of states, chains, and analysis techniques were detailed, and adaptations for modeling failure configurations in communication networks were presented. Additionally, algorithms for performing cumulative and up-to-absorption analyses were described. Besides that, a Markov reward cost model was introduced as a solution for modeling possible costs associated with states or transitions.

The addressed economic concepts provided a general overview of cost modeling for communication networks and detailing specific financial evaluation metrics such as TCO, CAPEX, OPEX, RoI, and NPV.

Finally, the presented content in this chapter forms a solid foundation for readers to understand the proposed frameworks in the later chapters.

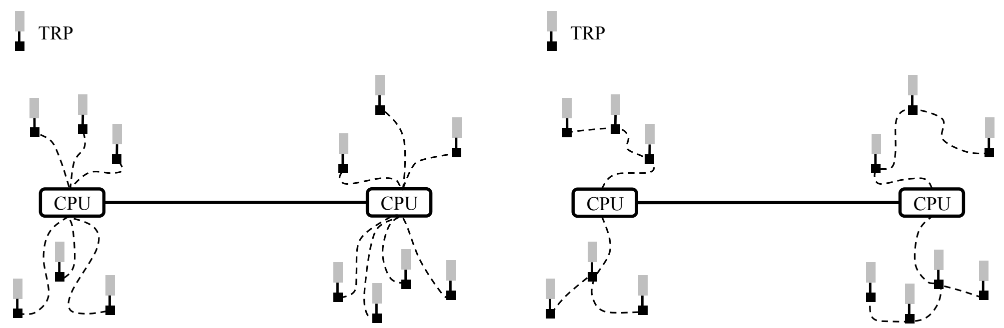
4 Reliability evaluation framework for cell-free mMIMO: failure impacts and fronthaul protection schemes

This chapter introduces a framework based on MCMC to evaluate the reliability aspects of cell-free mMIMO networks with segmented fronthaul, contemplating the first hypothesis and proposal of this thesis, described in Subsection 1.3.1. The provided analysis aims to assess the impact of both individual and cumulative failures on TRPs and fronthaul segments. Furthermore, it explores alternatives to mitigate these impacts if necessary, ultimately ensuring the feasibility of deploying cell-free mMIMO networks with segmented fronthaul.

The segmentation procedure involves connecting TRPs serially in a compute-and-forward architecture, as illustrated in Figure 28(a). This approach contrasts with the typical cell-free mMIMO configuration, which generally employs a star topology, featuring separate links between each TRP and a CPU, as depicted in Figure 28(b). In this context, the segmented fronthaul can significantly reduce the deployment complexity and enhance network scalability (Interdonato et al., 2019)(Shaik; Björnson; Larsson, 2020). Despite this, there is a risk of potential high-impact hardware failures due to serialization. Hence why, individual and cumulative failure analyzes are necessary.

Figure 28 – Illustrations of fronthaul topologies for cell-free mMIMO networks considered in the literature.

(a) Star. (b) Segmented fronthaul.



Source: elaborated by the author. .

A failure in a fronthaul segment or TRP causes sequential transport communication failure, i.e., fronthaul data outage in all the following TRP and fronthaul segments. This problem harmfully impacts macro-diversity and, consequently, SE, reducing the technical feasibility of cell-free mMIMO with segmented fronthaul. A typical approach to compensate for outage effects in communication networks is the utilization of protection/redundancy schemes. Some schemes exist for fixed access networks and cellular mobile networks (Selim et al., 2016) (Selim; Kamal, 2018)(Fernandez; Stol, 2016). Despite this, these schemes were not designed for the unique communication paradigms of cell-free networks and may require significant adaptations.

Finally, another important consideration is that segmented fronthaul can be implemented in a non-integrated manner, i.e., using individual equipment such as TRPs, cables, and transceiver interfaces, or in an integrated fashion, where all these components are part of a singular product. The ERSS exemplifies the latter approach, featuring circuit-mounted chips called Transmission-Reception Point Units (TRPUs) acting as TRPs serially connected within a cable or stripe using a shared bus. This setup provides power, synchronization, and fronthaul communication through the bus, which adopts a broadcast structure in the downlink and a pipeline structure in the uplink. Integrated solutions like the ERSS offer a cost-effective option for deploying cell-free mMIMO networks, as each stripe or cable requires only a single plug-and-play connection to the CPU. This setup facilitates network rollout in a true sense without requiring highly qualified personnel (Interdonato et al., 2019)(Shaik; Björnson; Larsson, 2020). Despite these advantages, integrated solutions still face potential reliability challenges from the segmentation procedure and may require protective measures within their integrated structures.

The subsequent sections present the MCMC reliability evaluation framework. It is used to analyze both integrated and non-integrated cell-free mMIMO networks, the former being represented by the ERSS. The network performance calculations in terms of user and fronthaul rates use the system model presented in Subsection 2.3. Finally, protection schemes for both non-integrated and integrated are proposed and evaluated, aiming to determine the best options for long-term operation in indoor environments.

4.1 MCMC reliability evaluation framework

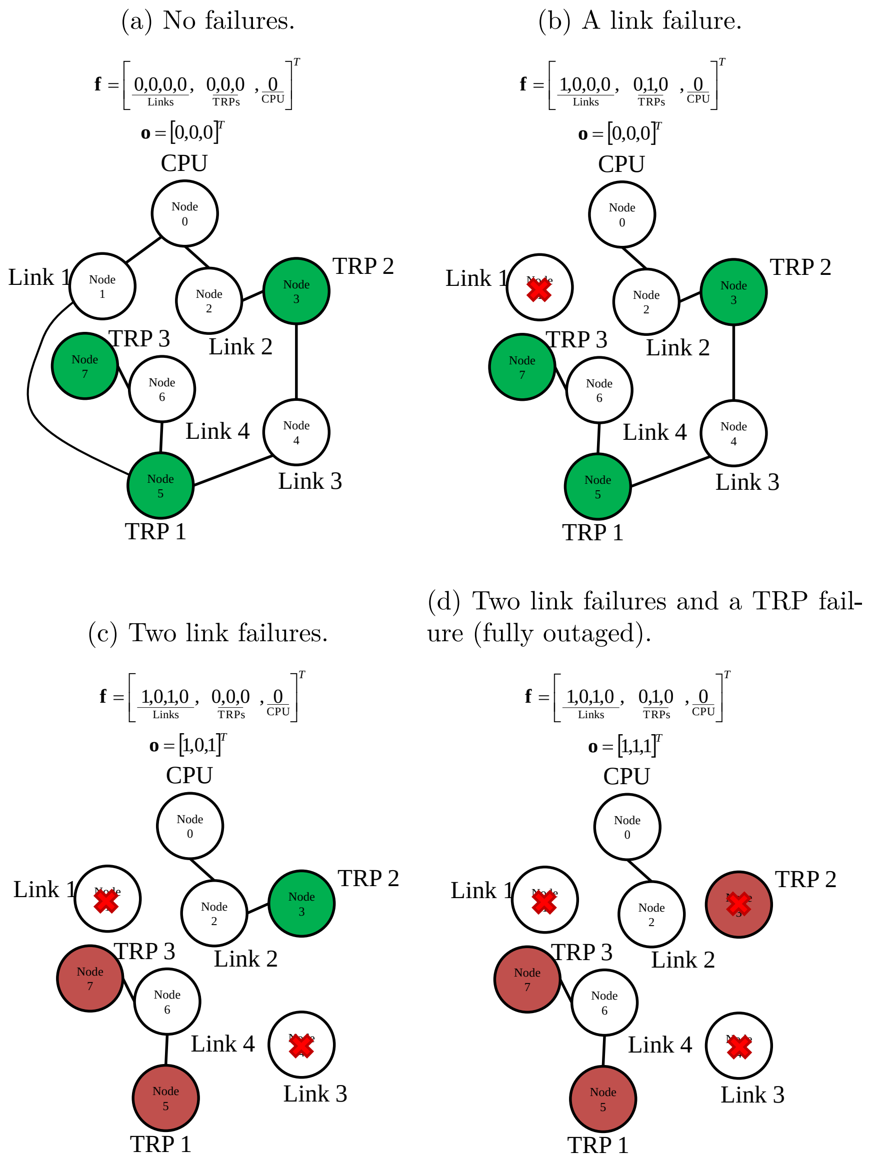
A cell-free mMIMO network with segmented fronthaul has a different total equipment count (Tot_{Eq}) depending on the fronthaul topology, medium, serialization level, and protection scheme. The same is true for an integrated cell-free mMIMO system but with components instead of equipment.

An undirected graph is utilized to represent the interconnections between network components. The symmetrical binary adjacency matrix of this graph is defined as $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, ..., \mathbf{A}_i, ..., \mathbf{A}_{Tot_{Eq}}]^T$, where $\mathbf{A}_i = [A_{i,1}, A_{i,2}, ..., A_{i,j}, ..., A_{i,Tot_{Eq}}]^T$. If equipment i is directly connected to equipment j, then $A_{i,j} = 1$. In this model, links are also equipment and, consequently, nodes due to the reasons described in Subsection 3.1.4. The $\mathbf{A}$ matrix models the connections of a fully functional network, and to account for failures it is modified to $\mathbf{A}^{failed}$ using a binary array of failures $\mathbf{f} = [f_1, f_2, ..., f_i, ..., f_{Tot_{Eq}}]$, where 1 indicates a failure on equipment i. Essentially, if f_i is equal to 1, $\mathbf{A}_i^{failed} = \mathbf{0}_{Tot_{Eq}}$, otherwise, $\mathbf{A}_i^{failed} = \mathbf{A}_i$.

The set of equipment directly connected to the CPU is given by $\mathcal{P}$. If a component i is inside $\mathcal{P}$, then i is a source node. This consideration is essential to verify if an TRP is m on SB l is adequately connected. In this case, its node must be achievable from at least one source node considering the modified adjacency matrix $\mathbf{A}^{failed}$. Then, to represent disconnection in the fronthaul links or TRPs due to component failure a binary outage array $\mathbf{o} = [o_{1,1}, o_{1,2}, \cdots, o_{1,M}, \cdots, o_{l,m}, \cdots, o_{L,M}]$ is considered. In other words, when $o_{l,m}$ is equal to one, the node of TRP m on SB l is not achievable any source node in $\mathcal{P}$.

Figure 29 exemplifies the formation of $\mathbf{o}$ from $\mathbf{f}$. It presents a network graph with one CPU, three TRPs, and four links. In Figure 29(a), there are no equipment failures,

Figure 29 – Examples of the formation of the outage array from the failure array. The graph represents a network with four links, three TRPs, and one CPU, having 8 nodes and edges.



Source: elaborated by the author.

and all TRPs are functional, resulting in an outage array of zeros. In Figure 29(b), a link failure occurs, modifying the adjacency matrix to represent the failure and changing the edges on the graph. Despite this, all TRPs can still reach the CPU, so the outage array remains an array of zeros. In Figure 29(c), another link fails, again modifying the adjacency matrix. This time, it is clear that the nodes equivalent to TRPs 1 and 3 can no longer reach the CPU node, being under outage, resulting in $\mathbf{o} = [1, 0, 1]^T$. Finally, in Figure 29(d) a failure occurs at TRP 2, altering the adjacency matrix and edges once more. With TRP 2 unable to reach the CPU node, the network becomes fully outaged, i.e., $\mathbf{o} = [1, 1, 1]^T$.

The user rate calculation under hardware failures needs to disregard TRPs under outage. This can be done by modifying the diagonal matrix $\mathbf{D}_{l,m,k} \in \mathbb{N}^{N\times N}$ presented in Section 2.3.2 to

$$\mathbf{D}^{\text{modified}}_{l,m,k} = \begin{cases} \mathbf{D}_{l,m,k} & \text{if } o_{l,m} = 0 \\ \mathbf{0}_{N_{l,m}} & \text{if } o_{l,m} = 1. \end{cases} \tag{4.1}$$

this modified matrix is then used in replacement of the normal $\mathbf{D}_{l,m,k}$ in Equations (2.42) to (2.45)

The impacts of instantaneous individual or even simultaneous failures of equipment/components can be evaluated by modifying $\mathbf{f}$. Despite this, the impacts of cumulative failures over time cannot be done by simply modifying some of the presented variables. To this end, a CTMC with state definition given by the different possible combinations of equipment/component failures ,i.e., $\Omega = \{1, 2, 3, ..., 2^{Tot_{eq}}\}$is considered. For example, a system with two TRPs connected by a fronthaul segment have the following possible failures: (i) none, (ii) TRP 1, (iii) segment, (iv) TRP 2, (v) TRP 1 and TRP 2, (vi) TRP 1 and segment, (vii) TRP 2 and segment, and (viii) all equipment (Anders; Silva, 2000).

The transitions rates are calculated as presented in Section 3.2 and cumulative analysis of the chain is performed. Due to the a large number of states, a MCMC simulation is used for this analysis. To this end, it is noteworthy that the probability of going from state i to state j ($P_{i,j}$) and the time of permanence in state i (T_i) before a transition are given by

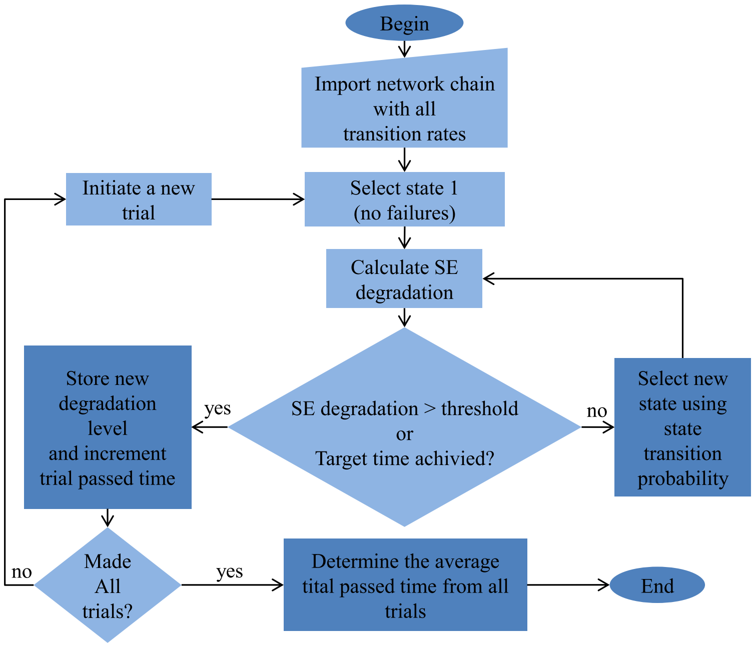
$$P_{i,j} = \frac{\lambda_{i,j}}{-\lambda_{i,i}}, \quad T_i = \tfrac{1}{-\lambda_{i,i}} \ln U, \tag{4.2}$$

where U denotes a random variable with a uniform distribution (Anders; Silva, 2000).

While the cumulative analysis is adequate to verify the degradation over time, it is inappropriate to determine the time until a specific SE degradation is achieved due to failures. An absorptive analysis should carries out this investigation. To this end, absorptive states are flagged as the failed ones with SE degradation under a pre-defined threshold, compared with the fully functional state.

Figure 30 presents the flowchart for the failure simulation procedure for both cumulative and absorptive analysis. For a pure cumulative analysis the SE threshold is higher than 1. For a pure absorptive the target time is infinity.

Figure 30 – Flowchart of the failure simulation procedure. The possible failures are modeled as a CTMC, and a cumulative or up to absorption analysis is carried out by a Monte Carlo simulation.



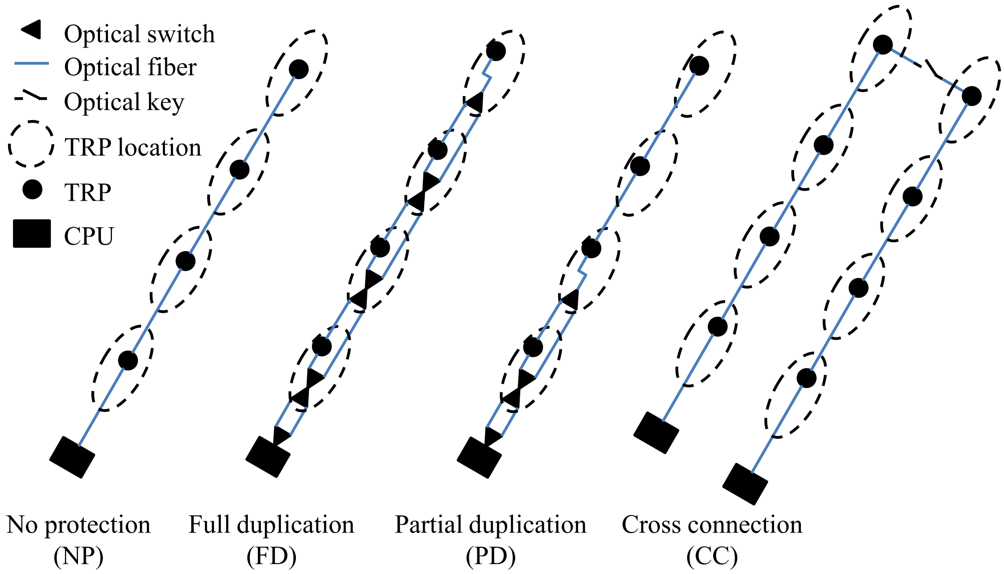
Source: elaborated by the author.

4.1.1 Protection strategies for non-integrated segmented cell-free mMIMO

The protection strategies to mitigate hardware failures within the transport network and TRPs in non-integrated cell-free mMIMO systems with segmented fronthaul vary depending on of access medium and technology employed. Possible fronthaul technologies include microwave, fiber, and copper, each requiring specific adaptations to integrate the appropriate specialized switches, keys, and redundant links. Based on this knowledge, it is feasible to apply some protection strategies from fixed-access communication networks to cell-free networks. Despite this, it requires some modifications to fit the unique architecture cell-free.

Given the high data rates anticipated in the fronthaul of next-generation communication systems, a fiber-based fronthaul is commonly assumed. Traditional protection schemes for fiber access networks, such as Optical Line Terminal (OLT) cross-connections and fiber duplication, have been adapted for non-integrated cell-free mMIMO systems with segmented fronthaul. This adaptation differs markedly from those employed in fixed fiber access networks. For instance, the duplication strategy requires $2M_l + 1$ switches for M_l links in SB l, while the fully equivalent fixed access protection would use $2M_l$ switches

Figure 31 – Illustration of the considered fronthaul architectures of the cell-free mMIMO network with segmented fronthaul.



Source: elaborated by the author.

for M_l links in SB l. This additional switch is employed to create dual links outside and within each TRP site, with one of the links at each site designed to bypass the TRP. This ensures continued connectivity in case of a TRP failure. Moreover, the cross-connection strategy is implemented at the furthest end of the serial chain of connections from the CPU. This placement contrasts with fixed access setups, where cross-connections are typically proximal to the OLT, which is analogous to the CPU in cell-free.

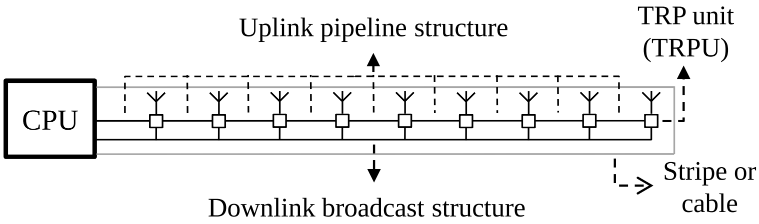
Figure 31 presents the considered protection schemes for a cell-free mMIMO network with fiber-based segment fronthaul, including: (i) Non-Protected (NP), (ii) Full Duplication (FD), (iii) Partial Duplication (PD), and (iv) Cross-Connection (CC). FD duplicates fronthaul SBs entirely and creates a bypass on-site TRP link to bypass possible TRP failure, i.e., TRPs are not duplicated. PD is similar to FD but only protects a percentage of the TRPs in the SBs length. The idea is to reduce costs by only duplicating more impactful links, similar to the duplication of only feeder fibers in fixed fiber access systems. Finally, CC creates cross-connections between fronthaul SBs.

4.1.2 Protection strategies developed for integrated segmented cell-free mMIMO

The ERSS is an integrated segmented cell-free mMIMO system. The TRPUs are serially connected to a CPUs using a shared fronthaul bus that provides power, synchronization, and fronthaul communication (broadcast structure for downlink and compute-and-forward for uplink), as illustrated in Figure 32 (Interdonato et al., 2019)(Shaik; Björnson; Larsson, 2020).

Protection solutions for ERSS will need to identify if a failure occurs in a bus or

Figure 32 – Illustration of the fronthaul structure of the ERSS, an integrated segmented cell-free mMIMO system.



Source: elaborated by the author.

a TRPUs. If one TRPU fails, the downlink broadcast structure and power will not be compromised, and only uplink fronthaul data from TRPUs after the failed TRPU will need to be rerouted. If a bus segment fails, everything after the failure (fronthaul data downlink/uplink and power) will need to be rerouted.

The failure verification is made by the "last" TRPU on the stripe (the ERSS SB, which is located farthest from the CPU in terms of the number of fronthaul segment hops. If no fronthaul is received on the bus downlink broadcast structure even after the "last" TRPU does a connection requisition with the CPU, it is a bus failure. If there is downlink fronthaul data but no acknowledgment of the data sent by the "last" TRPU to the CPU, then it is an TRPU failure disturbing the uplink pipeline structure.

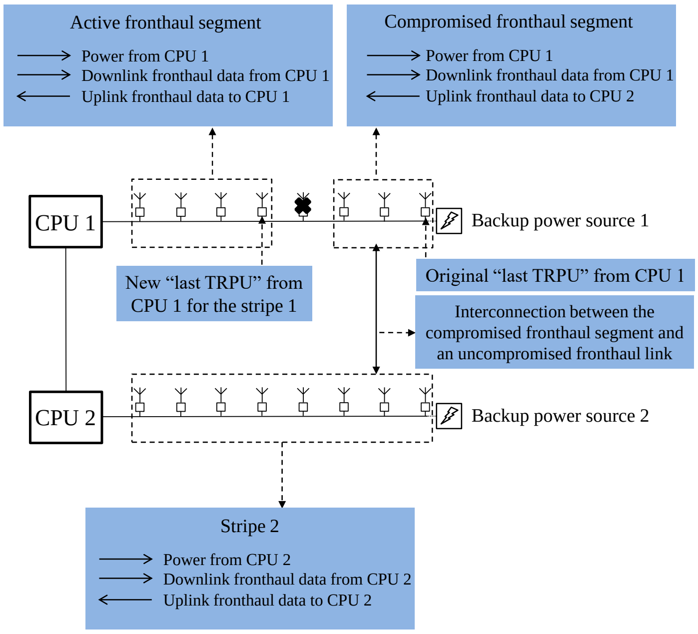
After identifying the type of failure, the last TRPU sendS the failure information through the uplink pipeline structure to other compromised TRPUs. It is important to remember that both bus segment and TRPU failures disrupt the uplink pipeline structure, and then failure information will never reach non-compromised TRPUs.

Then the compromised TRPU can reroute their fronthaul traffic through interconnections, although only uplink fronthaul data is rerouted for TRPU failure. As for the power loss due to bus failures, a backup power source connected to the last TRPU can be responsible for feeding power to the compromised TRPUs. Figure 33 exemplifies the failure identification and recovery procedure for TRPUs and bus failures. It is noticeable that a new "last" TRPUs is selected on the non-compromised part of the stripe. This is made to guarantee that the method can cover possible new failures.

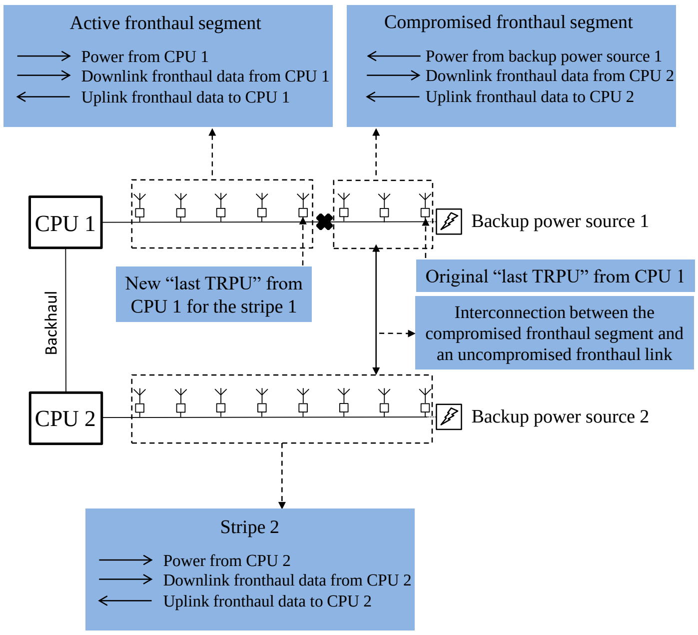
The fronthaul interconnection can be established using different technologies. Figure 34 provides a simplified overview of two alternatives: (a) Interconnection via redundancy fronthaul links and (b) Interconnection via TRPUs. The first approach attaches the redundancy fronthaul links through non-TRPU circuit-mounted chips (called switching units) in the regular fronthaul links. In contrast, the second uses wireless connections between TRPUs, which are established using dedicated wireless resources allocated for protection in each TRPU or group of TRPUs. If such resources are unavailable, the system

Figure 33 – A brief illustration of the proposed self-healing method for fronthaul communication failures in serial cell-free networks.

(a) Example of compensation for TRPU on serial chain failure.



(b) Example of compensation for fronthaul bus failure.



Source: elaborated by the author.

alternatively utilizes unused or low-loaded TRPUs.

In both wired and wireless cases, the fronthaul interconnection procedure is initiated by the stripe last TRPU, i.e., it is initiated in a distributed way, without CPU dependence. Besides that, multiple fronthaul interconnections technologies can be used simultaneously, e.g., interconnections via redundancy fronthaul links and TRPUs to be used concomitantly, which could be useful since it would provide a higher degree of failure recuperation with fewer redundancy fronthaul links.

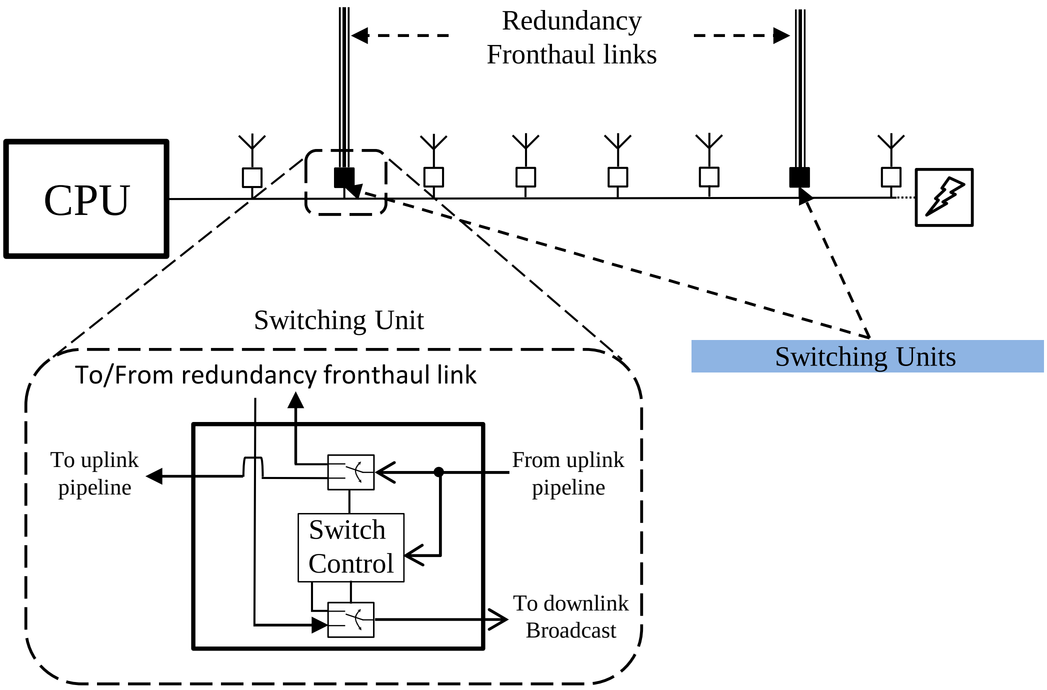
The developed protection method detailed procedure is presented in Figure 35. The presented method does failure identification and compensation in a distributed fashion. It requires little to no additional components for protection since it can be done aerially through unused or low-loaded TRPUs. The second power source is required only for the fronthaul bus failure protection, mainly in wireless interconnections, as the wired alternative can effectively transmit power. Nonetheless, the network designer can disregard bus failures if backup power hardware is not desired. This consideration can be done because TRPU failures are expected to be much more common than bus failures since transceivers, interfaces, and analog/digital communication equipment/components fail much more than the wired links (Berghmans; Eve; Held, 2007).

Further analysis of Figure 35 reveals that the method also considers coordination and negotiations between CPUs. This fact may seem strange since the method is presented as a distributed approach for failure recovery. However, this affirmation is used in the sense that the TRPUs identifies and initiates the failure recovery procedure. Once a TRPU requests an interconnection, the decision to support this request will involve the CPUs of the interconnecting stripes and is usually based on available resources and the optimization of the set of TRPs and stripes serving different users, seeking to reduce baseband and scheduling complexities or to ensure minimal latency.

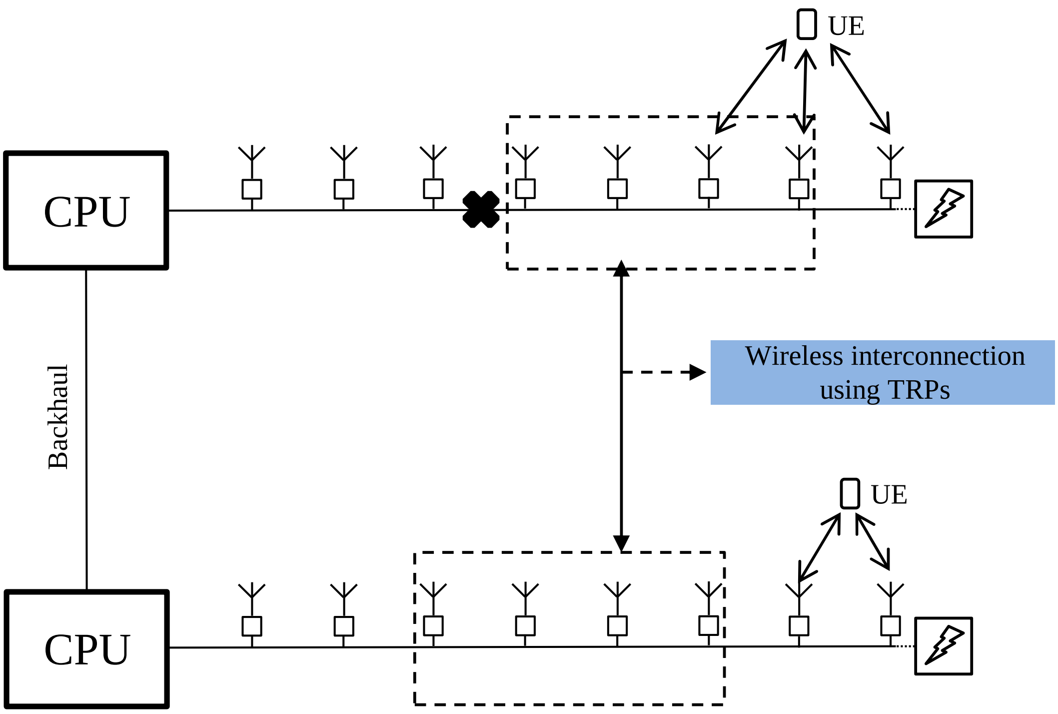
The compromised section of a stripe can request interconnections to multiple other stripes and even the uncompromised section of itself. In this way, the involved CPU may need to negotiate the best candidates for interconnection, considering latency, data provisioning, who perform main baseband operations and scheduling for the TRPUs on the compromised section of the stripe and the users they serve. The negotiations include the CPU of the compromised segment, especially when compensating for TRPU failures, as the affected segment's downlink data still passes through its CPU. Further details on fronthaul interconnections and resource negotiations within the ERSS can be found in Frenger et al. (2019).

Figure 34 – Simplified overview of two fronthaul interconnection technologies compatible with the method: (a) and (b) Wireless interconnection with TRPU specific resources or unused and low-loaded TRPUs.

(a) Wired interconnection.

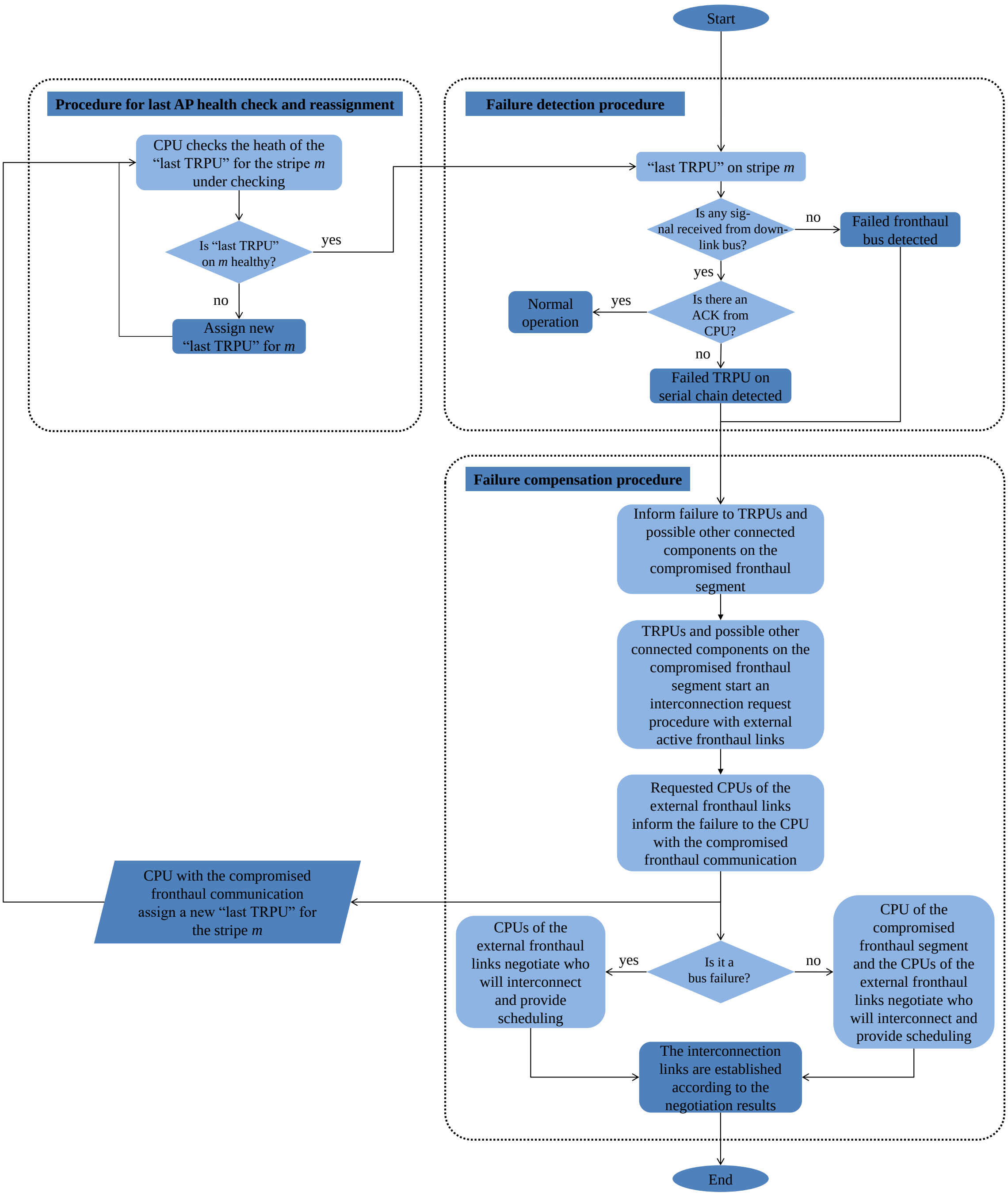


(b) Wireless interconnection.



Source: elaborated by the author.

Figure 35 – Flowchart for proposed fronthaul failure identification and compensation method for the ERSS.



Source: elaborated by the author.

4.2 Numerical results

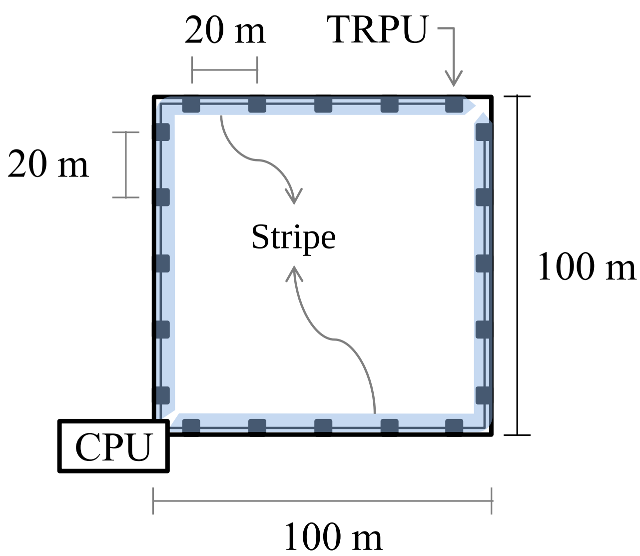
This section presents the results of this chapter. First, a proposed protection solution for a integrated cell-free mMIMO solution based on the ERSS is evaluated. Then non-integrated cell-free mMIMO solutions are also assessed. Both analysis consider an indoor scenario that can strongly can benefit from extended maintenance intervals due to protection schemes, which avoid noticeable performance degradation from cumulative failures in the perception of the users.

4.2.1 Integrated cell-free mMIMO

4.2.1.1 Case study

Figure 36 depicts the considered indoor scenario with an area of 100 m × 100 m, which operates under an ERSS with a TRPU spacing of 20 m. Two stripes (the ERSS SB) are deployed at the height of 5m along the walls, each with ten TRPUs.

Figure 36 – Illustration of the considered scenario for integrated cell-free mMIMO reliability results.



Source: elaborated by the author.

The failure compensation technology is based on the wireless interconnection detailed in Subsection 4.1.2, assuming that the wireless link possesses the capacity to fulfill the fronthaul requirements necessary to maintain functionality across the stripes. This can be achieved by either allocating separate resources at each TRPU explicitly for potential backup fronthaul usage or by utilizing currently unused or lightly-loaded TRPUs. The assumption of available fronthaul capacity is not unrealistic, given that the array of distributed antennas in fronthaul communication between stripes forms a MIMO link with high data rate capacity between them (Frenger et al., 2019).

The results that would be obtained for redundant wired fronthaul links may exhibit some differences, as failures would manifest slightly differently due to fixed possible interconnection positions, implying that additional TRPUs may become disconnected in each failure compensation event. However, the analysis of the proposed failure compensation strategy for the stripes with just the wireless interconnection can still be conducted without loss of generality.

The simulation parameters reflect a 3GPP Indoor Hotspot Open Office (InH-open) scenario operating in mid-band with low mobility UEs (3 km/h), which is compatible with indoor environments. The considered channel model was the uncorrelated Rayleigh fading, and there was no pilot contamination. Moreover, MR precoding was assumed in all TRPUs. These considerations are among the simplest ones in cell-free mMIMO systems but are sufficient to evaluate the proposed failure compensation strategy without loss of generality. In fact, the MR precoder and uncorrelated Rayleigh fading can be considered worst-case assumptions for a cell-free system, as correlated channels and more advanced precoders can be explored to cancel interference better (Björnson; Hoydis; Sanguinetti, 2017)(Demir; Björnson; Sanguinetti, 2021). Moreover, by not focusing on factors like pilot contamination, all performance loss becomes associated with the failures of TRPUs.

Finally, each TRPU aims to serve the four strongest UEs concerning itself and performs the heuristic power allocation presented in (Björnson; Sanguinetti, 2020b). This method allocates power to individual users by normalizing their large-scale fading gains by the summation of large-scale fading gains of all users within each TRPU. Moreover, the failure rate of each radio stripe was considered to be equivalent to the failure rate of an LED strip light (Luxalight, 2021), equivalent to a a failure rate of 1.42×10^{-6} per meter of strip. Additionally, it was considered that 80% of the failures happen at TRPUs. Table 4 shows the remaining parameters used in simulations.

4.2.1.2 Results

Figure 37 illustrates the Cumulative Distribution Function (CDF) versus UE SE for various potential individual failures in a system lacking any protection scheme. A "first

Table 4 – System, channel, and signal simulation parameters.

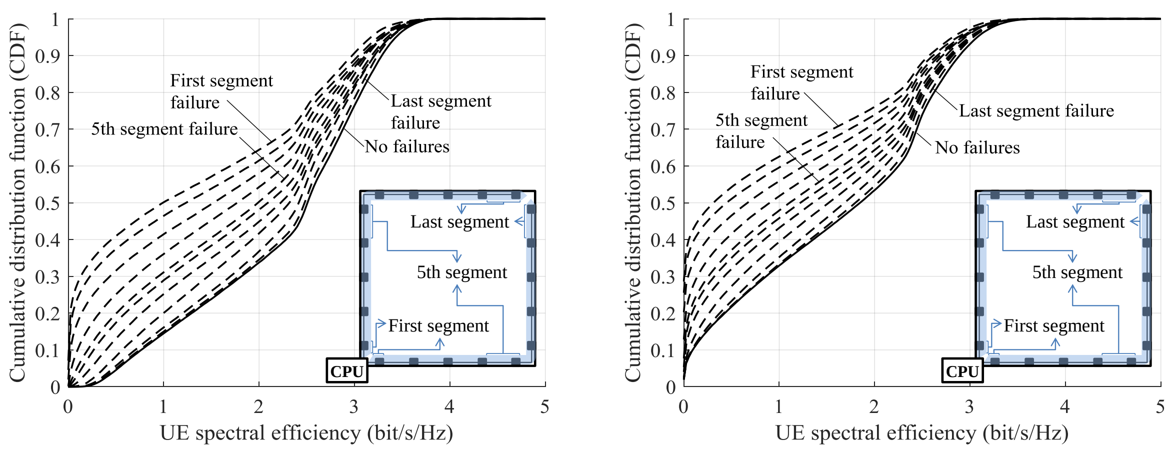
Parameter	**Values**
Number of UEs	8 or 16
Coherence block samples	3857
Bandwidth	100 MHz
TRPU total Tx power	12 dBm
TRPU antenna gain	3 dBm
UE total Tx power	22 dBm
Carrier frequency	3.5 GHz
Rx noise figure	8 dB

segment" failure denotes a fault in the fronthaul segment and/or TRPU directly connected to the CPU. Conversely, a "last segment" failure refers to faults in the final fronthaul segments and/or TRPU in the serial chain of connections, which are the farthest from the CPU. Notably, any failure can significantly affect all users' performance. For eight users at the 80th percentile, a representation of the best-performing users, a performance reduction of up to 15% in SE is observed, while at the 20th percentile, reflecting the worst-performing users, a reduction of up to 98% in SE is noted. For 16 users, the 80th percentile experiences a 15% reduction, and at the 20th percentile, the rate drops to zero.

The results in Figure 37 indicate that failures impact the worst-performing users substantially more than the best-performing ones. In both user scenarios, it is observed that failures up to the penultimate segments exhibit similar performance to scenarios with no failures. Additionally, failures in the first four segments are most detrimental, completely nullifying the rate for the 16-user case. Hence, the protection of any failure is beneficial, especially those at the first 40% length of the serial chain of connections.

Figure 37 – CDF versus UE SE for individual failures in a system without any protection scheme. The failures can happen at 10 possible segments in one stripe. Each segment encloses a TRPU and fronthaul connection.

(a) 8 users. (b) 16 users.



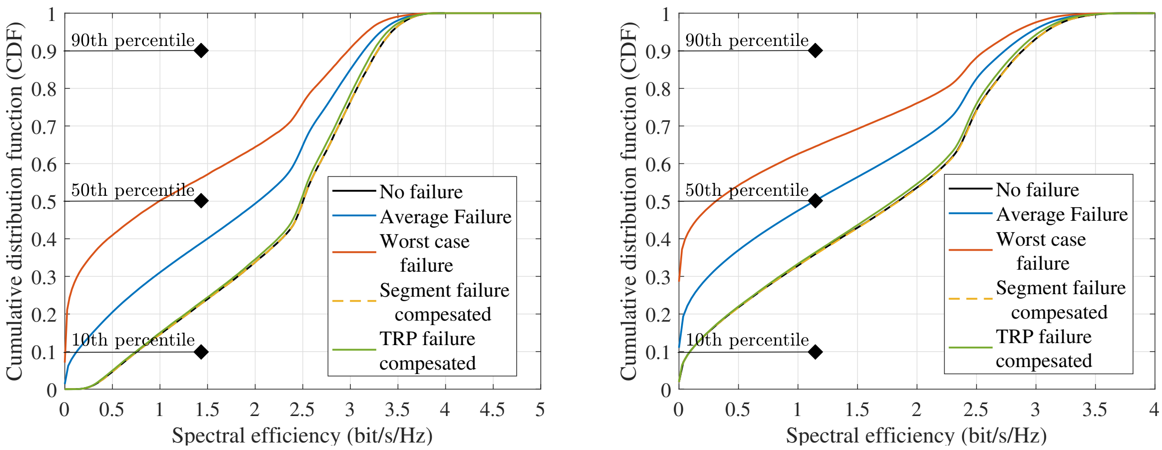
Source: elaborated by the author.

Figure 38 presents the CDF of UE SE in a system experiencing individual failures across five failure compensation scenarios: average failure case without compensation, worst failure case without compensation, compensated fronthaul bus failure, and compensated TRPU failure. All configurations are compared to a fully functional network. The average case is computed based on the SE of all possible segment failures, as shown in Figure 37 for a non-protected system, weighted by their likelihood, while the worst case corresponds to a failure in the first segment of a stripe. It is possible to note that configurations with failure compensation can almost completely mitigate the effects of failures. If an individual fronthaul segment failure occurs in any part of the stripe in a system with compensation, the SE performance is the same as the fully functional network.

Concerning TRPU and fronthaul bus failure compensation. Figure 38 results shows that the former has a slightly smaller SE than the latter, mainly because one TRPU is always lost. Despite this, the TRPU compensation still offers much better performance than the average failure in non-protected networks, which is equivalent to losing multiple TRPUs. In terms of numbers, considering the 50th percentile, an individual failure has, on average, 20% to 30% less impact in protected systems. The upper limit on this impact reduction was achieved in the more crowded 16-user scenario. Finally, in protected systems, the impacts are reduced by more than 80% concerning worst-case failures.

Figure 38 – CDF versus UE SE for individual failures under TRPU and fronthaul segment failure compensation in comparison with a non-protected network under the absence of failures, as well as under average and worst-case failures.

(a) 8 users. (b) 16 users.



Source: elaborated by the author.

Results until now highlight the substantial benefits of protection schemes in mitigating individual failures. Nevertheless, analysis of cumulative failures remains critical. In this regard, Table 5 provides an analysis of the average time to a 20% reduction in SE due to cumulative failures, both with and without the compensation method. The results demonstrate that the proposed protection method for a ERSS system can extend the duration before a 20% SE loss occurs by four times for the 8-user scenario and three times for the 16-user scenario.

This last result is important, as performing repairs after every failure can often be impractical. When minor failures occur, a few TRPUs may be disconnected, resulting in a performance degradation that, while measurable, may not be immediately perceptible in the high TRPU count cell-free environment. Therefore, substantial system repairs, such as replacing a stripe or conducting other forms of repair, are justified only when a noticeable performance loss occurs. So if the threshold for the performance loss is 20%, the protection schemes can quadruple the service life of unsupervised serial cell-free mMIMO networks based on ERSS for eight users.

Table 5 – Cumulative failures SE degradation analysis.

Failure compensation method employed?	**Number of Users**	**Average time until a 20% SE degradation due to cumulative failures (h)**	**Standard Deviation (h)**
No	8	4990	110
	16	3440	65
Yes	8	21535	270
	16	11000	218

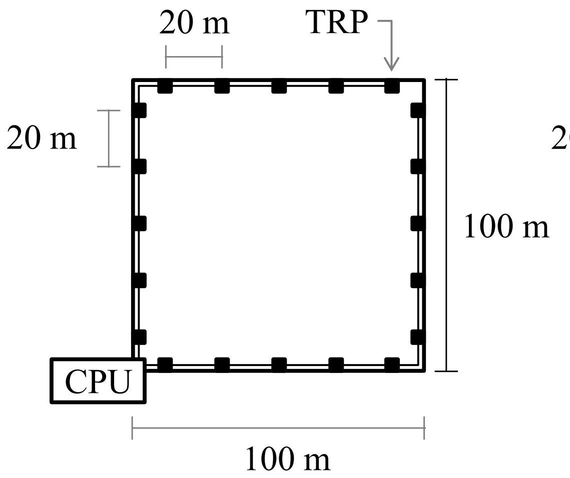
4.2.2 Non-integrated cell-free mMIMO

4.2.2.1 Case Study

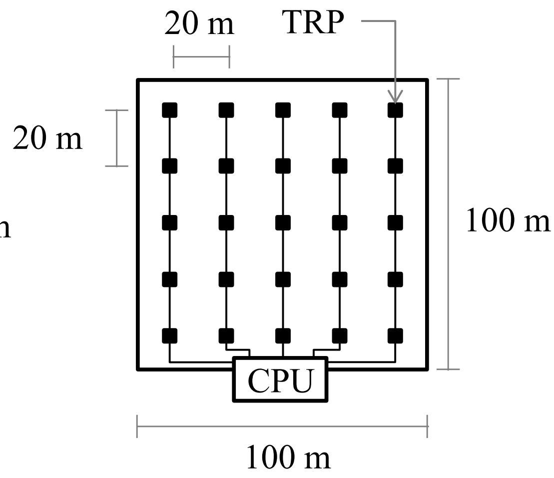
Figure 39 depicts the considered TRP configurations for the non-integrated cell-free mMIMO deployed in a indoor scenario with an area of 100 m × 100 m. In the first configuration, shown in Figure 39(a), the TRPs are distributed over the walls in two fronthaul SBs, whereas in the second one, shown in Figure 39(b), the TRPs are distributed over the area in five parallel fronthaul SBs. Additionally, each TRP has $N = 4$ antennas, optical fiber fronthaul connections, and is spaced from other TRPs by at least 20 m. Finally, UE and TRP heights are 1.65 m and 5.0 m, respectively.

Figure 39 – Illustration of the considered scenarios: (a) TRPs and fronthaul SBs in the walls along the area's perimeter, and (b) TRPs uniformly distributed in the ceiling serially connected by five parallel fronthaul SBs.

(a) Two SBs.



(b) Five SBs.



Source: elaborated by the author.

Each TRPs performs MR precoding and implements a heuristic power allocation method outlined in (Björnson; Sanguinetti, 2020b). This method allocates power to individual users by normalizing their large-scale fading gains by the summation of large-scale fading gains of all users within each TRP. The adoption of MR is due to its simplicity and to evaluate the reliability under the worst-case performance of cell-free mMIMO systems.

The simulation parameters reflect a 3GPP InH-open scenario operating in mid-band with low mobility UEs (3 km/h), which is compatible with indoor environments. This configuration resembles the one in Section 4.2, where it was observed that protection schemes can effectively mitigate failures in integrated cell-free systems. It is reasonable to assume that the same should hold for non-integrated cell-free systems. Therefore, a more realistic correlated Rician fading model is adopted to provide a more practical analysis.

In this model, the LoS probability for calculating the Rician factor is determined using probability equations from (3GPP, 2020) for the InH-open scenario. This approach is more intricate and closely aligns with practical cell-free implementations, where TRPs are situated near users and exhibit strong LoS channel components to the UEs, which are corrected within the antenna domain of each TRP. Consequently, the correlation matrices $\mathbf{R}_{l,m,k}$ adhere to the Gaussian local scattering model outlined in Subsection 2.3.1. Table 6 summarizes the remainder of simulated system parameters.

Table 6 – System, channel, and signal simulation parameters.

Parameter	**Values**
Number of UEs	8 or 16
Number of pilots	8
Maximum number of UE per TRP	8
Coherence block samples	3750
Bandwidth	100 MHz
TRP total Tx power	12 dBm
TRP antenna gain	3 dBm
UE total Tx power	22 dBm
Carrier frequency	3.5 GHz
Rx noise figure	8 dB
Angular standard deviation	$\sigma_\varphi = 20°$
Uniform linear array antenna spacing	half-wavelength

It is assumed a joint pilot assignment and TRP selection, where the first τ_p UEs are assigned mutually orthogonal pilots, and the remaining UEs are assigned to the pilot that experiences the lowest pilot contamination. Then, each TRP selects up to τ_p UEs with the highest average channel gain in each pilot (Björnson; Sanguinetti, 2020b). In this context, since 8 pilots are assumed, pilot contamination happens only when there are 16 UEs among the number of UEs assumed at Table 6.

For the failure simulations, four cases are considered: (i) NP, (ii) FD, (iii) PD, and (iv) CC, as illustrated in Figure 31. All of them are detailed in Subsection 4.1.1, in such a way that PD goes up 40% of the SBs length. This consideration is based on results for

Table 7 – Considered MTBF for the network different equipment.

Network element	**Aplication**	**MTBF (h)**
TRP	TRP	5.2×10^5 [a]
Fiber patch cord	Fronthaul link	10^8 (Berghmans; Eve; Held, 2007)
Optical switch or key	Protection: FD, PD, and CC	5×10^6 (Berghmans; Eve; Held, 2007)
Small Form-factor Pluggable (SFP)	TRP/Fronthaul interface	2.3×10^6 [b]

[a]Based on the MTBF of the following multi-user MIMO commercial TRPs series: Cisco aironet, Aruba 320, and Extremewireless.
[b]Based on Cisco SFP-10GSR.

individual failures from Subsection 4.2.1.2, where it was shown that failures in the initial 40% length of a non-protected integrated cell-free system have more impact on SE.

At last, the failure rates are calculated as indicated in Subsection 3.2 considering the MTBF values shown in Table 7.

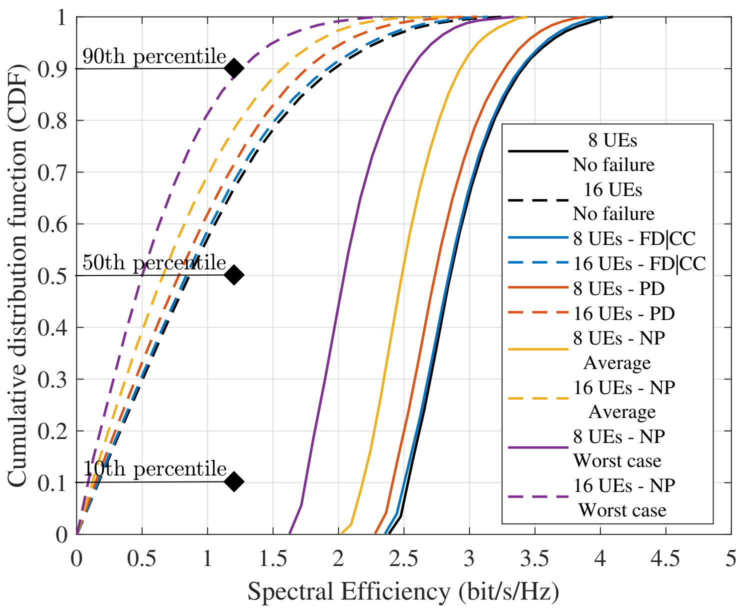
4.2.2.2 Results

Figure 40(a) presents the CDF of the SE under average individual failures for all considered protection schemes and NP case for the scenario with two SBs. One can understand this average individual failure SE as the mean of the SEs of all possible failures weighted by their probability of occurrence. Additionally, for comparison purposes, the Figure also shows the CDF of the SE of a fully functional network and the worst possible individual failure. The curves in the Figure prove that all protection schemes can significantly reduce the impacts of an individual failure, with CC and FD obtaining results very close to a fully functional network. For the non-contaminated 8 UE cases, failures similarly impact worst and best UEs. In contrast, for contaminated 16 UE cases, the impacts of failures are more evident on best UEs since the SE of the worst UEs is significantly reduced due to the lower interference cancellation arising from degraded channel estimates. Taking the 50th percentile region as a baseline, the average SE reduction due to failures in NP configurations is 23% for 16 UEs and 13% for 8 UEs.

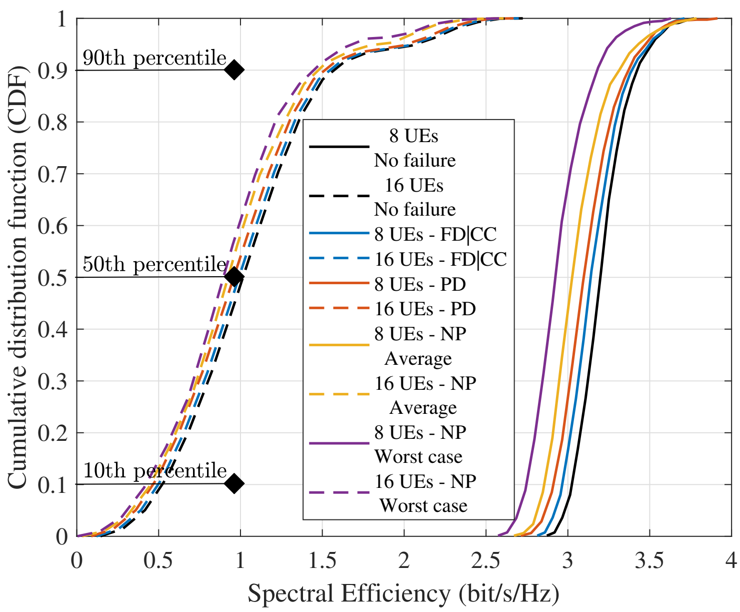
Figure 40(b) presents the CDF of the SE under average individual failures for all considered protection schemes and NP case for the scenario with five SBs. As in the previous figure, for comparison purposes, the figure also shows the CDF of the SE of a fully functional network and the worst possible individual failure. Once again, The results demonstrate that all protection schemes can mitigate the impacts of individual failures. Notably, the impact of individual failures is smaller than previous results, indicating that

Figure 40 – Comparison of the CDFs of the SE per UE concerning fully functional and failed networks with single faulty equipment across different protection strategies. The failed network performance is based on the average SE of all possible equipment failures weighted by their probability of occurrence. An exception is the worst-case curves, which illustrate the SE of the worst possible individual failure.

(a) Figure 39(a) scenario.



(b) Figure 39(b) scenario.



Source: elaborated by the author.

reducing the number of serially connected TRPs by half substantially increased the system reliability against individual hardware failures. Nevertheless, protection mechanisms can still mitigate a potential 10% reduction in SE from the worst-case failure to merely 2%. Moreover, an important observation is that while worst-case individual failures have less impact, they occur more frequently as the system has more SBs. This, coupled with cumulative failure analysis, can justify the adoption of protection schemes even in deployments with fewer TRPs per SB.

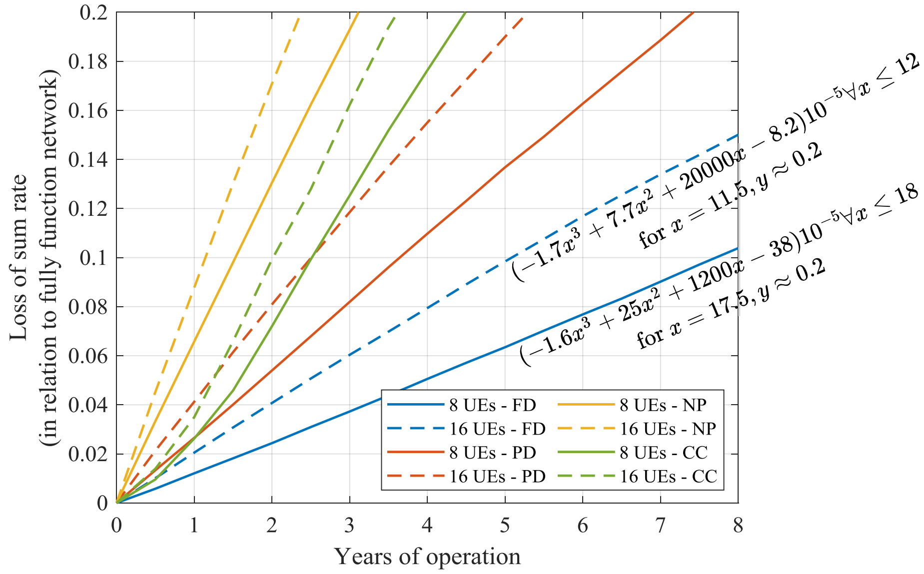
Figure 41 illustrates the sum rate degradation over time due to cumulative failures for all considered protection schemes and the NP case. The analysis is confined to a 20% degradation and is presented for the scenario involving two SBs. As can be seen, protected schemes can delay a 20% SE degradation by 1.1 to 14.4 years compared to the NP case. FD exhibits the lowest degradation over time, offering greater reliability for both considered user counts. Furthermore, PD demonstrates the second-best time to reach the maximum admitted degradation.

Despite this, Figure 41 reveal that for the two SBs scenario, CC than PD exhibits lower degradation until 1 and 1.4 years of operation for 8 and 16 UEs, respectively. However, beyond these points, the presence of multiple failures increases the degradation of CC significantly , way more than PD, which is expected as CC does not contribute much in terms of redundant equipment. Nevertheless, the fact that PD had 2.6 to 3 times more time until degradation compared to CC under the NP scenario indicates that the advantage of PD is much stronger in the considered scenario. This fact, together with the lower additional redundant equipment of PD concerning FD suggests that PD may be the best implementation candidate for the two SB scenario, as it can be less cost-prohibitive than FD.

Figure 42 depicts the sum rate degradation over time due to cumulative failures for all considered protection schemes and the NP case. The analysis is limited to a 20% degradation and is presented for the scenario involving five SBs. The results demonstrate a longer time until a 20% SE degradation for all protection approaches compared to the scenario with two SBs. In such a way that the curves underscore the advantages of the protection schemes for this scenario more than the SE CDF results, as the time until a 20% SE degradation due to cumulative failures is at least four years higher for a protection scheme compared to NP. Another interesting observation is that the five SBs scenario is well suited for CC, which exhibits the second-lowest degradation for 16 users and even eight users until 8.25 years of operation. This fact indicates that the presence of more alternative interconnection paths in a scenario with more SBs compensates for the lack of redundant equipment in CC, making it more feasible, especially since it can be very cost-effective due to its minimal redundancy equipment.

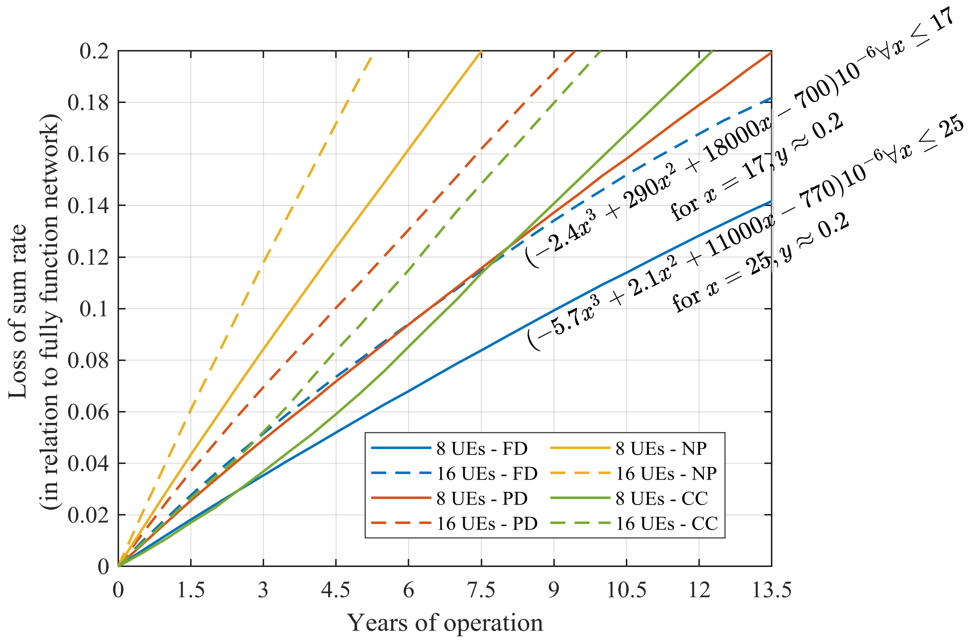
Another way to interpret the results of Figs. 41 and 42 is to consider programmed

Figure 41 – Average time until a sum rate degradation of 0 to 20% due to cumulative failures for the different considered fronthaul protection architectures in the two SBs scenario.



Source: elaborated by the author.

Figure 42 – Average time until a sum rate degradation of 0 to 20% due to cumulative failures for the different considered fronthaul protection architectures in the five SBs scenario.



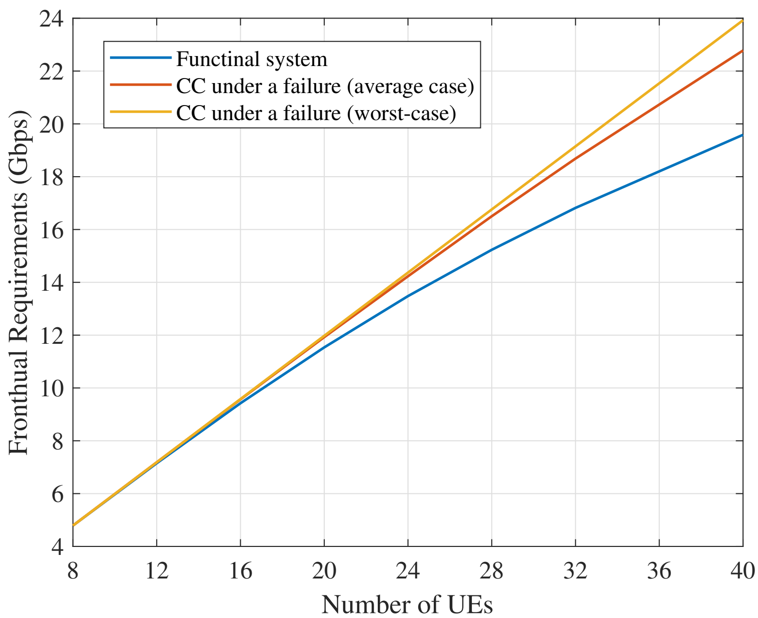
Source: elaborated by the author.

maintenance. For example, for the two SBs scenarios, 8 UEs and yearly maintenance, sum rate degradation is 6.5%, 2.5%, 2.5%, and 1% for NP, CC, PD, and FD, respectively. If maintenance is made when the yearly degradation of NP is achieved, CC, PD, and FD would only need maintenance every 1.8, 2.3, and 5.2 years. The reduced amount of maintenance can justify the upfront cost of the protection schemes' implementation.

A potential challenge in the CC scheme is the risk of fronthaul network overload due to the transfer of TRPs between SBs. Figure,43 evaluates this issue by illustrating the expected maximum fronthaul requirements for scenario (a) in Figure,39, considering a quantization with 3 bits per sample, similar to configurations in Femenias and Riera-Palou (2020). The Figure,43 presents curves for the fully functional case, the average number of TRPs transferred to another SB when a failure occurs, and the worst-case scenario, where all TRPs are transferred to the same SB. By analyzing the figure, it is clear that with the original assumption of up to 16 users, a 10 Gbps Ethernet fronthaul interface would be adequate to support both failed and non-failed configurations. Besides that, notable differences in the fronthaul requirements between fully functional and failed scenarios only occur for more than 16 users, with the differences becoming more pronounced from 20 users onwards. For 40 users, the increase in fronthaul requirements is 16% for the average case and 22% for the worst case. However, this user count is an extreme case since the ratio of total TRP antennas to users is 2.

In scenarios with fewer serially interconnected TRP, the impact of fronthaul network overload should be even less pronounced, as less TRP are transferred to another SB upon a failure. Thus, while the fronthaul design may need careful consideration to support the CC scheme, the additional requirements are not substantial enough to be a major bottleneck for the scheme's feasibility. In many cases, systems could probably implement CC without any need for additional fronthaul capacity.

Figure 43 – Fronthaul bandwidth requirements under CC protection for scenario (a) in Figure39.

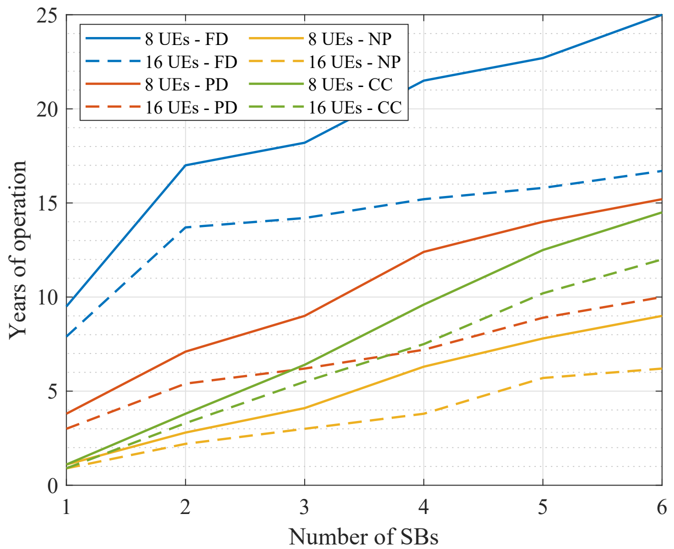


Source: elaborated by the author.

4.2.2.3 Sensitivity analysis

The results indicate that reliability improves when L increases and M decreases. To investigate this more deeply, Figure 44 presents a sensitivity analysis verifying the time until a 20% degradation in the sum rate for a scenario with 100 m × 100 m, 24 TRPs, and different numbers of SBs. The NP (with more than 2 SBs, CC, and PD results are approximately linear. In a way that the impacts of failures double each time the number of SBs doubles. Besides, at the beginning of the curves, time to 20% degradation in sum rate is longer in PD than CC. However, the curves of CC grow faster with the number of SB than PD. In this way, the first surpasses the latter in 4 SBs for 16 UEs, and if the trend holds, it should surpass in 7 SBs for 8 UEs. These results indicate that CC is more desirable when the number of SBs is at least 4, and the number of TRPs per SB is less or equal to 6.

Figure 44 – Sensitivity analysis of the time until a 20% sum rate degradation varying number of SBs in a 100 m × 100 m scenario with 24 TRPs. Solid lines represent 8 UEs and dashed lines represent 16 UEs. The curves in green, yellow, blue, and magenta represent FD, CC, PD, and NP, respectively.



Source: elaborated by the author.

To provide an idea of the costs associated with the protection strategies, Table 8 presents the expected number of deployed devices for all considered protection strategies, assuming the previously described indoor scenario with 24 TRPs connected to the CPU through 2, 4, and 6 SBs. The CAPEX variation concerning the schemes will occur due to the additional redundancy equipment in the protection strategies. The OPEX variation is mainly expected to be related to repair costs. Energy costs will remain similar across all schemes due to the use of passive equipment and similar processing and fronthaul demands. The latter holds even for CC, as illustrated in Figure 43.

For the described scenario, CAPEX can be streamlined by assuming equal costs

Table 8 – Expected number of deployed devices for all considered protection strategies (24 TRPs).

Equipment	**NP**	**FD**	**PD**	**CC (2 SBs)**	**CC (4 SBs)**	**CC (6 SBs)**
Links	24	48	34	26	30	34
Simple switches	0	48	20	0	0	0
Simple keys	0	0	0	1	3	5
Total	24	96	54	27	33	39

for deploying and acquiring links, simple switches, and keys, considering them equivalent to 1 Cost Unit (CU). Other types of equipment are present in equal quantities across all strategies and may not significantly impact the analysis. The OPEX related to repair is more nuanced because in a cell-free mMIMO network, where users can connect to multiple TRPs, repairs are only necessary when a noticeable performance degradation occurs. This approach allows multiple failures to be addressed simultaneously, with repair costs involving equipment replacement and extended work hours for fault identification, replacement, and testing. To simplify the OPEX, it can be assumed that all repair procedures cost half of the CAPEX of the unprotected base case. This assumption is reasonable for the considered small indoor network scenario, where large repairs might require a comparable timeframe to the initial network deployment.

Based on these assumptions, Table 9 presents figures for CAPEX, OPEX, and TCO over a 20-year operational period for all protection strategies with 2, 4, and 6 SBs. Repairs were initiated upon a 20% degradation in SE due to cumulative failures. It is evident that for two SBs, the lowest TCO was achieved with partial duplication, and even full duplication resulted in lower costs compared to the NP case. For 4 SBs, CC emerges as a more feasible option, a trend that continues with 6 SBs. An interesting observation is the significant TCO reduction observed from 2 to 6 SBs in the NP case, primarily due to its lower CAPEX and enhanced reliability from fewer TRPs connected in series.

Table 9 – Expected number of deployed devices for all considered protection strategies (24 TRPs).

Protection Type	**CAPEX**			**OPEX(Repair)**			**TCO**		
	2 SBs	*4 SBs*	*6 SBs*	*2 SBs*	*4 SBs*	*6 SBs*	*2 SBs*	*4 SBs*	*6 SBs*
NP	24	24	24	109.1	63,2	40	133.1	87.2	64
FD	96	96	96	17.5	15,8	14.4	113,5	111.8	110.4
PD	54	54	54	44.4	33,3	24	98.4	87.3	78
CC	27	33	39	72.7	32	20	99.7	65	59

Based on the strong cost-decrease behavior of NP when the number of SBs increases, one might consider having one fronthaul link for each TRP as a more cost-effective solution, effectively abandoning serial connection between TRPs. However, it is essential to recognize that the presented analysis is highly simplified and serves only as a basic cost comparison among protection approaches. Implementing a full star topology fronthaul network could introduce complexities and higher costs due to increased link quantities and longer installed link lengths. This can potentially impact fronthaul capacities, failure rates, and other critical parameters depending on the network size and scenario. These variables were not considered in the table analysis, where it can be noted that CAPEX remained consistent from 2 to 6 SBs in the NP case.

4.3 Chapter summary

This chapter introduced a reliability framework for assessing isolated and cumulative hardware failures in cell-free mMIMO networks with segmented fronthaul. In this context, the study presented in this chapter contemplated the analysis for the first proposal and hypothesis of the thesis, detailed in Subsection 1.3.1. The framework was based on graph theory and MCMC simulations, which were used to evaluate failure impacts and protection strategies in integrated and non-integrated cell-free mMIMO systems.

For integrated systems, it was considered an ERSS system and a proposed wireless or wired interconnection between stripes, which are the SBs that connect multiple TRPs in the ERSS. The proposed protection solution executed failure identification and compensation in a distributed manner, requiring minimal or no changes to TRP and ERSS hardware. For non-integrated systems, an optical-based fronthaul network was adopted, and three optical fronthaul architectures were proposed and examined: FD, PD, and CC, which were compared with an NP deployment.

Two case studies compared the SE performance of cell-free mMIMO under failures, with and without protection, in an indoor office scenario with 100 x 100 m relevant to the telecommunication industry. The results showed that the impacts of failures could not be neglected, particularly for users in the 50th percentile or lower. It was noted that failures in the initial 40% length of the serial chain of connections were more impactful. Based on this result, PD strategically duplicated the initial 40% of the fronthaul network on the SB.

The proposed protection solutions for the ERSS and non-integrated systems essentially mitigated the impacts of any individual failure on the system. Furthermore, they extended the time until a 20% SEdegradation due to cumulative failures when repairs were not executed by 1.5 to 4 times, depending on the considered system and protection scheme. This can potentially reduce the frequency and necessity of maintenance while extending the service life of unsupervised cell-free mMIMO networks.

Concerning the protection schemes for non-integrated systems, FD was the most reliable for minimizing SE degradation. However, PD has less installed equipment and good performance under cumulative failures. CC performed well under a higher count of SBs, starting at four in the provided analyses, which was equivalent to connecting six or fewer TRPs serially. A simplified cost analysis showed that CC is the most feasible option when four or more SBs are available. Before this point, PD was more affordable.

Considering all findings, it is evident that ensuring the feasibility of cell-free mMIMO networks with segmented fronthaul requires at least a configuration that includes a large number of SBs with fewer access points per segment. For instance, a 24 TRP system configured with 2 SBs will experience degradation three times faster than one configured with 6 SBs. Even then, significantly higher reliability can be achieved through protection alternatives, even for systems with fewer SB counts. This proves that protection schemes are essential for the feasibility of cell-free mMIMO networks with segmented fronthaul.

5 Techno-economic evaluation framework for cell-free mMIMO: A comparison between centralized and distributed processing

This chapter proposes a techno-economic framework designed to evaluate different types of deployments for cell-free mMIMO networks, contemplating this thesis's second hypothesis and proposal, described in Subsection 1.3.2. The framework includes a comprehensive cost assessment methodology for cell-free mMIMO networks, which addresses the TCO while accounting for deployment configurations, processing implementations, computational demands, and fronthaul signaling. To this end, existing literature models were adapted and integrated with newly developed models for deploying cell-free mMIMO systems and their components.

The cost assessment methodology explores two proposed TRP deployment strategies. The first strategy solely supports UE demands, while the second also aims to provide a fairer service distribution among UEs. The methodology further includes pricing and energy consumption models based on hardware computational and fronthaul requirements. The latter is determined by a model that calculates the necessary bitrate to maintain a specified level of SE degradation due to fronthaul quantization. Essentially, the model sets a minor degradation in UE SE concerning a non-quantized fronthaul as a target, establishing the minimum bitrate needed to achieve this degradation.

The proposed methodology quantifies all critical technical aspects of operating a cell-free mMIMO network, clarifying its capabilities and requirements and transforming it from a theoretical concept into a practical deployment that can be used to evaluate the feasibility of new solutions. This chapter uses the methodology to compare the feasibility of distributed and centralized processing implementations for cell-free mMIMO to determine which configuration is more viable under various conditions and constraints.

The architecture of cell-free mMIMO is inherently distributed due to the scattered TRPs that coherently serve users. However, baseband and processing functions can be executed either locally at the TRPs or at edge CPUs, allowing for both distributed and centralized implementations, which can be respectively understood as the BTRP and BCPU functional splits detailed in Subsection 2.3.3. In terms of characteristics, BCPU allows for more sophisticated techniques that enhance interference cancellation and SE, but it is computationally demanding. In contrast, BTRP aligns well with the distributed nature of cell-free mMIMO and consumes fewer computational resources due to its simpler processing techniques.

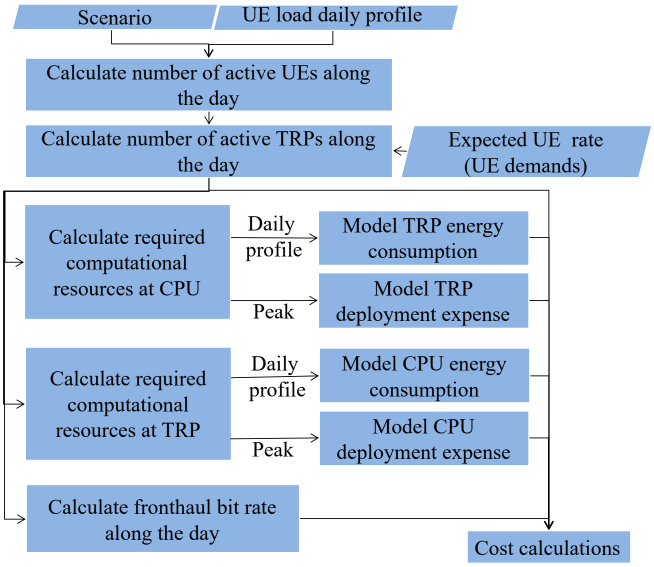
The debate over which type of processing implementation is more feasible is ongoing in the literature (Ngo et al., 2017)(Interdonato et al., 2019)(Björnson; Sanguinetti, 2020a)(Demir; Björnson; Sanguinetti, 2021). Initially, distributed processing was favored for its simplicity, which was believed to enhance system scalability and reduce fronthaul signaling requirements (Ngo et al., 2017)(Interdonato et al., 2019). However, as scalable processing techniques for centralized systems were developed and fronthaul signaling estimations showed that centralized processing might require less fronthaul signaling than distributed systems in the context of cell-free networks, the balance shifted in favor of centralized processing (Björnson; Sanguinetti, 2020a)(Björnson; Sanguinetti, 2020b). Nevertheless, this does not mean centralized approaches are always superior. The computational complexity in centralized systems can be orders of magnitude higher than in distributed setups (Freitas et al., 2023). Additionally, fronthaul requirements may increase in centralized scenarios, depending on factors such as the number of antennas on the TRP, the number of supported UEs, and the necessary sample bit width to accommodate the data rates of the supported UEs (Björnson; Sanguinetti, 2020a)(Femenias; Riera-Palou, 2020). Moreover, centralized processing lacks the general adaptability of distributed processing, which can also be implemented in a centralized fashion (Demir et al., 2024). In this context, a comprehensive techno-economic comparison is essential to adequately assess the superiority of centralized or distributed processing in different situations.

5.1 Cost assessment methodology

The proposed methodology to assess the total cost of a cell-free mMIMO is presented in Figure 45. It begins with a predefined scenario that includes propagation characteristics, the maximum number of UEs, and existing infrastructure. Moreover, a UE load daily profile characterizes the active UE ratio at different hours, while an expected UE rate represents UE demands.

These inputs drive calculations for the number of active UEs and TRPs along the day. The latter is chosen to support the expected UE rate in the provided scenario. Then, computational resource requirements for CPUs and TRPs are calculated, with peak requirements used to model deployment expenses and the daily variation used to calculate daily energy consumption in TRP and CPU. Simultaneously, the methodology determines the necessary fronthaul bit rate to accommodate fluctuating active UEs and TRPs under the expected UE rate. Ultimately, the fronthaul bit rate, TRP, and CPU models are used alongside the total number of active and inactive TRPs to calculate the comprehensive costs of deploying and operating a cell-free mMIMO system.

Figure 45 – Proposed cost assessment methodology of a cell-free mMIMO system.



Source: elaborated by the author.

When different precoders are considered, the methodology is fully executed for each of them, where the number of active TRPs to support the UE expected rate becomes the main driver in performance difference between the precoders.

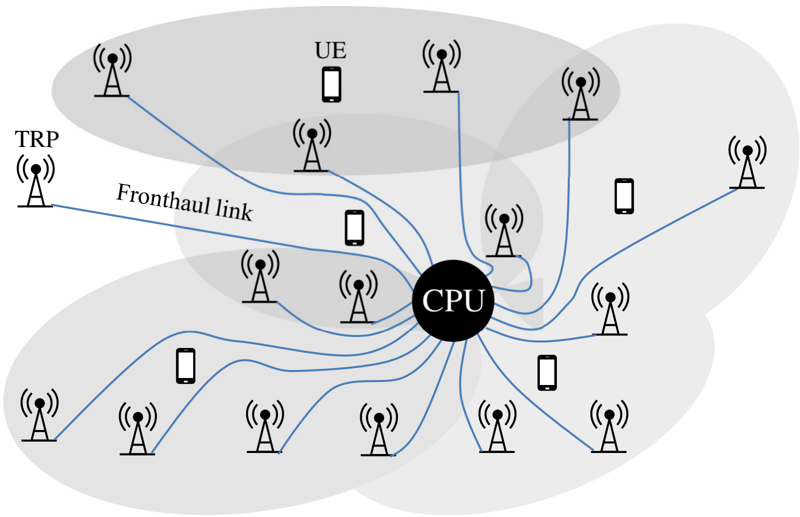
While the methodology can theoretically be applied to any number of TRPs and CPUs, this thesis specifically derives computational requirements for scenarios involving a single CPU, as outlined in Subsection 2.3.3.4. Analysis of multiple CPU scenarios is left for future work. In this context, this chapter assumes the presence of a single edge CPU supporting a limited coverage area with multiple TRPs, as depicted in Figure,46. Despite this, the case study presented in this chapter is well-suited for operations involving one edge CPU and has proven adequate for the analyses and comparisons undertaken.

The following subsections detail proposed TRP deployment strategies, outlining the number of necessary active TRPs required to support a given UE demand and potentially achieve a higher level of fairness. Additionally, calculations for the fronthaul bitrate with a non-fixed sample bit width are presented. The subsections also include structural models for edge CPUs and TRPs, which can provide estimates for the energy consumption and deployment costs of these device or infrastructure.

5.1.1 Number of active TRPs

A scalable cell-free mMIMO system with DCC ensures that each UE remains connected to at least one TRP (Demir; Björnson; Sanguinetti, 2021). In this scenario, the minimum viable count of TRPs is determined by the ratio between the number of UEs in the coverage area and the TRPs capacity in terms of UE connections. In this chapter, this capacity corresponds to the number of pilots (Björnson; Sanguinetti, 2020b). However, to

Figure 46 – Illustration of the system model considered network architecture in chapter 5. Dedicated fronthaul links connect a single edge cloud CPU to the TRPs. UEs are served by a limited optimal set of TRPs with available resources.



Source: elaborated by the author.

effectively enhance the capacity for UEs in a cell-free mMIMO system, it is desirable that the number of TRPs within the coverage area is much larger than the number of UEs present in that area (Interdonato et al., 2019).

These two constraints present two possible values for TRP count, one limited by coverage, i.e., restricted by the TRP maximum UE connections and coverage radius, and the other limited by capacity, i.e., to ensure the support of a given UE traffic demands requirement. In this context, the number of active TRPs inside a coverage area to support the UE load of the time t can be calculated similarly to as

$$M_{act,t} = \max\left(M_{act,t,C}, M_{act,t,R}\right), \tag{5.1}$$

where $M_{act,t,C}$ is the minimum number TRP to support all UEs inside the coverage area in time t, and $M_{act,t,R}$ denotes the number of TRPs necessary to provide the UEs with an expected rate R for the UE load of the time t (Fiorani et al., 2014).

Assuming that each individual TRP can have a effective communication channel to any UE in the entire coverage area, then $M_{act,t,C} = \rho_K \alpha_t\, S/\tau_p$, where ρ_K is UE density, α_t is the active UE load ratio at time t, and S symbolizes the coverage area. In other cases, the calculation of $M_{act,t,C}$ is more complex and not considered in this chapter, being left for future implementations[1].

There is no straightforward way to compute $M_{act,t,R}$. Nevertheless, obtaining an average rate equivalent to R for a given $M_{tot} = M_{act,t,R}$ and $K = \rho_K\, \alpha_t S$ is relatively simple

[1] The problem can be addressed by optimizing a clustering algorithm applied to UE positions, with the goal of minimizing the number of clusters. Constraints include a maximum cluster size of τ_p and the distance from a cluster element to its centroid not exceeding the TRP's maximum coverage radius. The variable $M_{act,t,C}$ is defined as the number of clusters.

using a Monte Carlo simulation process in conjunction with (2.56) (Demir; Björnson; Sanguinetti, 2021). In this context, it is possible to calculate a set of rates $\mathcal{R}$ for a specific set of TRP counts, defined by $\mathcal{M}_{tot} = \{M_{act,t,C}, M_{act,t,C}+M_{\text{step}}, M_{act,t,C}+2M_{\text{step}}, \cdots, M_{tot,\max}\}$, where $M_{tot,\max}$ is the maximum value of TRPs that can be implemented and M_{step} is the increment step for each element in $\mathcal{M}_{tot}$. This procedure results in $\mathcal{R} = \{R_1, R_2, \cdots, R_{|\mathcal{M}_{tot}|}\}$ where $R_1 < R_2 < \cdots < R_{|\mathcal{M}_{tot}|}$. Finally, the value for an arbitrary $M_{act,t,R}$ can be calculated using an interpolation process, which takes $\mathcal{M}_{tot}$ and $\mathcal{R}$ as inputs, as long as $R_1 < R < R_{|\mathcal{M}_{tot}|}$.

A rate R based on the average UEs' rates is a valid metric to evaluate the throughput of a communication system. However, this criteria can mask subtleties like rate variations between UEs under good and bad service quality, also called sometimes lucky and unlucky UEs. In this context, a R calculation based on a proportional fairness metric is proposed and used to perform a fairer TRP deployment actively. This way, both the basic average rate-based deployment and the proposed fairer one are used to provide a more thoughtful analysis of the network feasibility assessment.

The proposed fairer TRP deployment is established on a customer-based Service Level Agreement (SLA) with an agreed UE rate (Qureshi et al., 2021). Ensuring a fixed rate in mobile networks is challenging due to UEs' mobility and other random factors (El-Saleh et al., 2023). In this context, UEs may experience rates above or below the agreed rate. Nevertheless, the network ensures that at least a certain fraction of the agreed rate is consistently achieved, regardless of the UEs' disposition or location. This performance guarantee is denoted as a percentage of the agreed rate, represented by F_{rate}, which can vary between 0 % and 100 %. Additionally, this guarantee covers a portion of the coverage area, denoted by F_{cov} as a percentage ranging from 0 % to 100 %. This metric is labeled as SLA F_{rate}:F_{cov}.

From the CDF of achievable UE rates (Demir; Björnson; Sanguinetti, 2021), the agreed rate is calculated by

$$R_{\text{agreed}} = \min\left(R_{\text{acov}}, \frac{r_{F_{\text{cov}}}}{0.01F_{\text{rate}}}\right), \tag{5.2}$$

where $r_{F_{\text{cov}}}$ is the $(100 - F_{\text{cov}})$th percentile rate in the CDF and R_{acov} is the average rate of the achievable UE rates higher or equal to $r_{F_{\text{cov}}}$.

The expected UE rate for an SLA F_{rate}:F_{cov} TRP deployment is calculated as

$$R = \frac{100 - F_{\text{cov}}}{100} R_{\text{bcov}} + \frac{F_{\text{cov}}}{100} R_{\text{agreed}}, \tag{5.3}$$

where R_{bcov} denotes the average rate of the achievable UE rates smaller than $r_{F_{\text{cov}}}$. It is noticeable that the expected rate for the UEs with achievable rates larger than $r_{F_{\text{cov}}}$ is assumed to be the SLA agreed rate.

Finally it is important to mention that the deployed number of General Purpose Processors (GPPs) is calculated by $N_{\text{TRP}} = \sup_t M_{act,t}$.

5.1.2 Non-fixed bit width fronthaul bitrate calculation

The fronthaul bitrate of cell-free mMIMO networks, based on the BTRP and BCPU split implementations, was detailed in Subsection 2.3.3.3. It was shown that the primary factors influencing the bitrate include the antenna count per TRP, the number of users served by each TRP, the total available bandwidth of the radio signal, and the bit width used for samples of data and pilots.

In the literature, the bit width is usually pre-fixed or calculated for a given fronthaul capacity (Bashar et al., 2018)(Femenias; Riera-Palou, 2020)(Bashar et al., 2021). Depite this, the aimed cost-analysis nature of this thesis must allow different fronthaul capabilities at distinct costs. In this context, fixing the fronthaul capacity or the number of bits representing each scalar is undesirable. In this context, this chapter proposes the utilization of a maximum acceptable SE degradation due to quantized fronthaul samples parameter in bps/Hz (a_{deg}) to calculate the number of bits to represent the transmitted scalars. This approach allows the fronthaul bit rate to be associated with the theoretical UE rate performance. If a_{deg} is small enough, the network provides its best performance in terms of throughput.

Under a simplification where the same bit width is applied at a TRP level, in such a way that $b_{l,m'}^{\text{data}}$ and $b_{l,m'}^{\text{pil}}$ will remain constant for any $l \in 1, ..., L$ and $m' \in 1, ..., M_l$, as will $b_{l,k}^{\text{data}}$ for each $l \in 1, ..., L$ concerning equation 2.23. Algorithms 3 and 4 obtain the number of bits for the quantized data samples in the BTRP and BCPU splits, respectively. Both algorithms ensure that the SE degradation caused by fronthaul quantization does not exceed a_{deg}, even for the UE with the highest degradation. Besides that, the BTRP algorithm increments the bit width on a per-UE basis while trying to maximize the network throughput.

5.1.3 Required computational complexity capacity in CPU and TRPs

At different times t the system will have a different $M_{act,t}$, leading to changes in the required computational capacity, measured in GOPS, for edge CPUs and TRPs. These requirements can be calculated as described in Subsection 2.3.3.4. The fluctuations in computational requirements significantly affect both the cost and power consumption of edge CPUs and TRPs, as will be detailed in the upcoming subsections.

Algorithm 3: Bit Allocation Evaluation in a BTRP split when the same bit width is applied at a TRP level ($b^{\text{data}}_{l,k} = b^{\text{data}}_{l',k} \forall l' \in \{1, ..., L\}$).

Input: K, a_{deg}

bits $\leftarrow [1]_{1\times K}$ ▷ Initializes an array of ones representing the number of bits used to represent each UE's signal
SE $\leftarrow$ CALCULATE_SE(**bits**) ▷ Calculates the SE of each UE according to the number of bits used to represent each UE's signal
SE_target $\leftarrow$ CALCULATE_SE($\infty *$ **bits**) ▷ Calculates the SE of each UE for a fronthaul with unlimited capacity
while $\max_{i=1,\cdots,K}(\boldsymbol{SE\_target}[i] - \boldsymbol{SE}[i]) > a_{\text{deg}}$ **do**
  bits_crt $\leftarrow$ **bits**
  for *each* i ***in*** $\{1, 2, \ldots, K\}$ **do**
    bits_fut $\leftarrow$ **bits_crt**
    bits_fut$[i]$ $\leftarrow$ **bits_fut**$[i] + 1$
    SE_fut $\leftarrow$ CALCULATE_SE(**bits_fut**)
    if $(\boldsymbol{SE\_target}[i] - \boldsymbol{SE\_fut}[i]) > a_{\text{deg}}$ **then**
      if $\left(\sum_{j=1}^{K} \boldsymbol{SE\_fut}[j] > \sum_{j=1}^{K} \boldsymbol{SE}[j]\right)$ **then**
        SE $\leftarrow$ **SE_fut**
        bits $\leftarrow$ **bits_fut**
      end
    end
  end
end

Output: **bits** $= \{b^{\text{data}}_{l,1}, \cdots b^{\text{data}}_{l,K}\}$

Algorithm 4: Bit Allocation Evaluation in a BCPU split when the same bit width is applied at a TRP level ($b^{\text{data}}_{l,m'} = b^{\text{data}}_{l',m''} \forall l' \in \{1, ..., L\}$ and $m'' \in \{1, ..., M\}$).

Input: K, a_{deg}

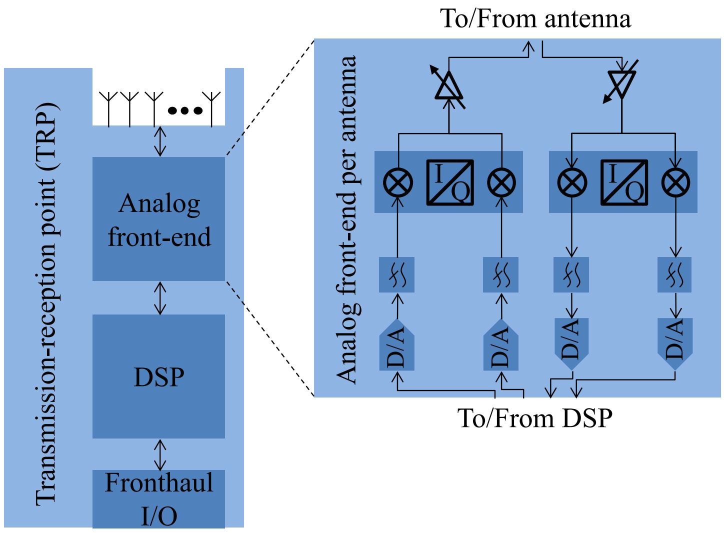
$b^{\text{data}}_{l} \leftarrow 1$
SE $\leftarrow$ CALCULATE_SE($b^{\text{data}}_{l,m'}$) ▷ Calculates the SE of each UE according to the bit width
SE_target $\leftarrow$ CALCULATE_SE(∞) ▷ Calculates the SE of each UE for a fronthaul with unlimited capacity
while $\max_{i=1,\cdots,K}(\boldsymbol{SE\_target}[i] - \boldsymbol{SE}[i]) > a_{\text{deg}}$ **do**
  $b^{\text{data}}_{l,m'} \leftarrow b^{\text{data}}_{l,m'} + 1$
  SE $\leftarrow$ CALCULATE_SE(b^{data}_{l})
end

Output: $b^{\text{data}}_{l,m'}$

5.1.4 TRP structure model

The TRPs are deployed throughout the coverage area to ensure effective communication between the network and UE devices. Besides the proper spacing between TRPs to improve coverage and signal distribution, it is important to model their components for power and cost modeling. Figure 47 provides an illustrative overview of the components of a non-vendor specific TRP in the cell-free mMIMO system. These components include antennas for bidirectional signal transmission, an analog front-end for initial radio signal processing, Digital Signal Processorss (DSPs) for tasks such as channel estimation and FFT/IFFTs conversions, and an Input/Outputs (I/Os) interface that facilitates seamless network communication (Interdonato et al., 2019; Demir et al., 2024; EARTH, 2010).

Figure 47 – Example of the components for a non-vendor specific TRP in a cell-free mMIMO system.



Source: elaborated by the author.

The analog front end comprises several subcomponents, such as Variable Gain Amplifiers (VGAs), In-Phase and Quadrature (IQ) modulators, filters, Digital-to-Analog Converter (DAC), and Analog-to-Digital Converter (ADC) converters. Figure 47 illustrates how they are interconnected for each antenna in the TRP. The subcomponents work together to adjust signal amplitudes, manage phase and frequency, refine bandwidth, and facilitate digital-to-analog conversion. They are usually designed to operate in synergy and can be integrated into a unified System on Chip (SoC) configuration (Interdonato et al., 2019; Demir et al., 2024; EARTH, 2010).

The power consumption of a TRP is influenced by its transmission power, accounting for losses during amplification, as well as the power usage of its individual components (Demir et al., 2024; Debaillie; Desset; Louagie, 2015; EARTH, 2010). In this context, it can be calculated as

$$P_{\mathrm{TRP}} = \mathrm{pw}_{\mathrm{DSP}} + \mathrm{pw}_{\mathrm{AFend}} + \mathrm{pw}_{\mathrm{IOint}} + \alpha_{\mathrm{amp}}\mathrm{pw}_{\mathrm{Tx}}, \tag{5.4}$$

where $\mathrm{pw}_{\mathrm{IOint}}$, $\mathrm{pw}_{\mathrm{DSP}}$, $\mathrm{pw}_{\mathrm{AFend}}$ are of the power consumption of I/O interfaces, DSP, and analog front-end, respectively. Besides that, $\mathrm{pw}_{\mathrm{Tx}}$ and α_{amp} represent the transmission power and an expansion factor to account for losses in the amplification process, respectively. The DSP power consumption is dependent on the computational complexity of the digital processing functions executed at the TRP, being calculated as

$$\mathrm{pw}_{\mathrm{DSP}} = \gamma_{\mathrm{pwDcore}} \left\lceil \frac{\mathrm{GOPS}_{t,R}^{\mathrm{TRP}_{l,m}}}{\mathrm{CAP}_{\mathrm{Dcore}}} \right\rceil + \gamma_{\mathrm{pwDSP}}\, \mathrm{GOPS}_{t,R}^{\mathrm{TRP}_{l,m}} + \mathrm{pw}_{\mathrm{DSP}}^{\mathrm{other}}, \tag{5.5}$$

where $\gamma_{\mathrm{pwDcore}}$ and γ_{pwDSP} are power slopes related to the DSP idle core operation and the number of operations in all cores, respectively. The variables $\mathrm{CAP}_{\mathrm{Dcore}}$ and $\mathrm{pw}_{\mathrm{DSP}}^{\mathrm{other}}$ are the GOPS capacity of a DSP processing core and a constant term representing other types of power consumption in the DSP, respectively. $\mathrm{GOPS}_{t,R}^{\mathrm{TRP}_{l,m}}$ denotes the required processing capacity in GOPS at TRP m on SBl to efficiently handle the network's UE load at a specific time t, while maintaining a data transmission rate of R, respectively. The analog front-end power consumption is given by

$$\mathrm{pw}_{\mathrm{AFend}} = 2N \left(2\mathrm{pw}_{\mathrm{filter}}^{\mathrm{ana}} + \mathrm{pw}_{\mathrm{IQmod}} + \mathrm{pw}_{\mathrm{VGA}} + \mathrm{pw}_{\mathrm{DAC}} + \mathrm{pw}_{\mathrm{ADC}} \right), \tag{5.6}$$

where $\mathrm{pw}_{\mathrm{filter}}^{\mathrm{ana}}$, $\mathrm{pw}_{\mathrm{IQmod}}$, $\mathrm{pw}_{\mathrm{VGA}}$, $\mathrm{pw}_{\mathrm{DAC}}$, $\mathrm{pw}_{\mathrm{ADC}}$ are the power consumption of analog filter, IQ modulator, VGA, DAC and ADC, respectively.

Similarly to power consumption, the price of the TRP can also be modeled by the individual prices of its components, being calculated as

$$\mathrm{pr}_{\mathrm{TRP}} = \mathrm{pr}_{\mathrm{DSP}} + \mathrm{pr}_{\mathrm{AFend}} + \mathrm{pr}_{\mathrm{IOint}} + N\mathrm{pr}_{\mathrm{ant}}, \tag{5.7}$$

where $\mathrm{pr}_{\mathrm{DSP}}$, $\mathrm{pr}_{\mathrm{AFend}}$, $\mathrm{pr}_{\mathrm{IOint}}$, $\mathrm{pr}_{\mathrm{ant}}$ are the prices of DSP, analog front-end, I/O interface and antennas, respectively. The price of the used DSP can be calculated as

$$\mathrm{pr}_{\mathrm{DSP}} = \gamma_{\mathrm{prDcore}} \left\lceil \frac{\mathrm{GOPS}_{\mathrm{peak},R}^{\mathrm{TRP}}}{\mathrm{CAP}_{\mathrm{Dcore}}} \right\rceil + \mathrm{pr}_{\mathrm{DSP}}^{\mathrm{base}}, \tag{5.8}$$

where $\gamma_{\mathrm{prDcore}}$ is a price slope for the necessary number of cores in the DSP, $\mathrm{pr}_{\mathrm{DSP}}^{\mathrm{base}}$ is a fixed price related to other DSP construction parameters, and $\mathrm{GOPS}_{\mathrm{peak},R}^{\mathrm{TRP}}$ is the peak required GOPS at a TRP to provide an expected UE rate R, obtained as the maximum of the set of supremums of $\mathrm{GOPS}_{t,R}^{\mathrm{TRP}_{l,m}}$ concerning t for each $l \in 1, ..., L$ and $m \in 1, ..., M_l$. The analog front-end price is given by

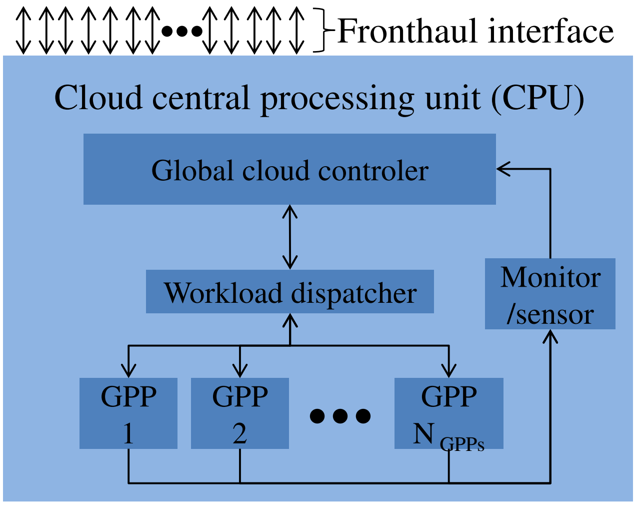
$$\mathrm{pr}_{\mathrm{AFend}} = \alpha_{\mathrm{SoC}} 2N \left(2\mathrm{pr}_{\mathrm{filter}}^{\mathrm{ana}} + \mathrm{pr}_{\mathrm{IQmod}} + \mathrm{pr}_{\mathrm{VGA}} + 2\mathrm{pr}_{\mathrm{DAC|ADC}} \right), \tag{5.9}$$

where $\mathrm{pr}_{\mathrm{filter}}^{\mathrm{ana}}$, $\mathrm{pr}_{\mathrm{IQmod}}$, $\mathrm{pr}_{\mathrm{VGA}}$, $\mathrm{pr}_{\mathrm{DAC|ADC}}$ are the prices of analog filter, IQ modulator, VGA, and DAC or ADC, respectively. Besides that, α_{SoC} is a price reduction factor due to SoC integration.

5.1.5 Edge CPU structure model

It is assumed that CPU functionalities are deployed virtually in one edge-cloud server, following the C-RAN workload consolidation model outlined in (Sigwele et al., 2017). The edge cloud CPU is then composed of Global Cloud Controller (GCC), workload dispatcher, GPPs, and monitor/sensors, as presented in Figure 48. The GCC converts UE traffic into manageable workloads and makes resource management. It ensures that the number of active GPPs aligns with the current workload, optimizing GPP utilization. The workload dispatcher distributes the workload among the GPPs, which executes the workload processing. The monitors/sensors collect utilization status from the GPPs and gather utilization data from the GPPs and transmit it back to the GCC. This information assists in proper workload management and resource allocation.

Figure 48 – Illustration of the edge cloud CPU workload consolidation model (Sigwele et al., 2017).



Source: elaborated by the author.

The workload capacity at a time t in the edge cloud CPU is given by the number of active GPPs, which is calculated as

$$N_{\mathrm{GPPs},t}^{\mathrm{act}} = \left\lceil \frac{\mathrm{GOPS}_{t,R}^{\mathrm{CPU}}}{\mathrm{CAP}_{\mathrm{GPP}}} \right\rceil, \tag{5.10}$$

where $\mathrm{CAP}_{\mathrm{GPP}}$ represents the capacity of the GPP in GOPS, and $\mathrm{GOPS}_{t,R}^{\mathrm{CPU}}$ denotes the required single CPU processing capacity in GOPS to efficiently handle the network's UE load at a specific time t, while maintaining a data transmission rate of R. The deployed number of GPPs is calculated by $N_{\mathrm{GPPs}} = \sup_t N_{\mathrm{GPPs},t}^{\mathrm{act}}$.

The GPPs are assumed to be housed in racks, each with a specific housing capacity. If the number of GPPs exceeds the capacity of a single rack, additional ones will be utilized. In this context, the space occupied by the edge cloud CPU depends on the number of racks and is given by

$$s_{\mathrm{CPU}} = \left\lceil \frac{N_{\mathrm{GPPs}}}{\mathrm{CAP}_{\mathrm{rack}}} \right\rceil s_{\mathrm{rack}}, \tag{5.11}$$

where $\mathrm{CAP}_{\mathrm{rack}}$ represent the maximum amount of GPPs that a rack can hold and s_{rack} is the necessary area to install a rack in m^2, which is larger than the area of the rack since extra space exists for equipment installation/maintenance, movement of personnel, and ventilation needs.

The power consumption of the entire edge-cloud CPU at a time t is calculated as

$$P_{\text{CPU},t} = P^{\text{IT}}_{\text{CPU},t} + P^{\text{cool}}_{\text{CPU},t}, \tag{5.12}$$

where $P^{\text{IT}}_{\text{CPU},t}$ is the power of IT components at time t, i.e., servers and network equipment, and $P^{\text{cool}}_{\text{CPU},t}$ is the power of the cooling system at time t (Hardy et al., 2013; Cui et al., 2017).

The power of the IT components at the time t is given by

$$P^{\text{IT}}_{\text{CPU},t} = \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil \text{pw}^{\text{rack}}_{\text{Net}} + P_{\text{GPP},t} N^{\text{act}}_{\text{GPPs},t}, \tag{5.13}$$

where $\text{pw}^{\text{rack}}_{\text{Net}}$ and $P_{\text{GPP},t}$ represent the power consumption of the network equipment per rack and the power consumption of the GPP at a time t, respectively. The latter component is calculated by

$$P_{\text{GPP},t} = \text{pw}^{\text{idle}}_{\text{GPP}} + \left(\text{pw}^{\text{peak}}_{\text{GPP}} - \text{pw}^{\text{idle}}_{\text{GPP}}\right) \frac{\text{GOPS}^{\text{CPU}}_{t,R}}{\text{CAP}_{\text{GPP}}\, N^{\text{act}}_{\text{GPPs},t}}, \tag{5.14}$$

where $\text{pw}^{\text{idle}}_{\text{GPP}}$ and $\text{pw}^{\text{peak}}_{\text{GPP}}$ are the idle and peak power consumption of the GPP, respectively (Hardy et al., 2013; Cui et al., 2017).

The cooling requirements of a server room mainly depend on its floor area and the heat generated by the IT and other electric equipment. The calculation of the requirements may be complex and require special software (Patterson; Costello; Grimm, 2007). Consequently, the power consumption of the cooling system in data centers can also be complex to calculate (Zhang; Liu, 2022). Despite this, if the cooling Power Usage Effectiveness (PUE) is known, the power consumption of the cooling system can then at a time t be calculated as

$$P^{\text{cool}}_{\text{CPU},t} = (\text{PUE}_{\text{cool}} - 1) P^{\text{IT}}_{\text{CPU},t}, \tag{5.15}$$

where PUE_{cool} is the PUE of the cooling system (Cui et al., 2017).

The pricing of the support infrastructure for the edge cloud CPU is calculated as

$$\text{pr}^{\text{CPU}}_{\text{Sinf}} = \left\lceil \frac{P^{\text{peak}}_{\text{CPU}} T_{\text{Pout}}}{\text{CAP}_{\text{bat}}} \right\rceil \text{pr}_{\text{bat}} + P^{\text{peak}}_{\text{CPU}} (\gamma_{\text{Co|PD}} + \gamma_{\text{inv}}) + \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil \text{pr}_{\text{rk\&nt}}, \tag{5.16}$$

where $P^{\text{peak}}_{\text{CPU}}$ is the edge cloud CPU peak power consumption, achieved when all deployed GPPs are fully active and utilized. Moreover, CAP_{bat}, T_{Pout}, and pr_{bat} are variables linked to the installed battery bank. Specifically, CAP_{bat} is the battery capacity in Wh factoring in the depth of discharge, T_{Pout} is the maximum duration of a power outage that can be managed, and pr_{bat} represent the cost for the battery's acquisition and installation.

Besides that, $\gamma_{\text{Co|PD}}$ and γ_{inv} stand for price slopes. The former indicates the cooling and power distribution infrastructure expense per Watt, while the latter pertains to the inverter costs of the backup power source per Watt. Finally, $\text{pr}_{\text{rk\&nt}}$ defines the acquisition and installation cost of a rack and the network equipment on a per-rack basis (Hardy et al., 2013; Jahid et al., 2020).

5.2 Cost models

This section presents the cost model utilized to determine the TCO of the cell-free mMIMO system in the context of the methodology in Figure 45. The model is divided into CAPEX and OPEX, which are summed to obtain the TCO. In this context, the CAPEX is given by

$$\text{CAPEX} = C_{a\&i}^{\text{CPU}} + C_{a\&i}^{\text{TRPs}} + C_{a\&i}^{\text{Xhaul}}, \tag{5.17}$$

where $C_{a\&i}^{\text{CPU}}$, $C_{a\&i}^{\text{TRPs}}$, $C_{a\&i}^{\text{Xhaul}}$ represents the acquisition and installation cost for CPU, TRPs, and fronthaul interfaces, respectively. Conversely, the OPEX is given by

$$\text{OPEX} = T_{\text{ope}}^{\text{hours}} \left(C_{\text{fSpace}}^{\text{hourly}} + C_{\text{rep}}^{\text{hourly}} + \frac{\text{pr}_{\text{kWh}}}{24} \sum_{n=1}^{N_{\text{samples}}^{\text{daily}}} P_n^{\text{total}} T_n^{\text{sample}} \right), \tag{5.18}$$

where $T_{\text{ope}}^{\text{hours}}$ is the adopted operational time in hours, pr_{kWh} is the price of Kilowatt Hour (kWh), $N_{\text{samples}}^{\text{daily}}$ is the considered number of time samples in a 24-hour period for the UE load variation, P_n^{total} is the total power consumption at each time sample n, and T_n^{sample} is the duration of each time sample n in hours, i.e., n is a discretization of t. Additionally, $C_{\text{fSpace}}^{\text{hourly}}$ and $C_{\text{rep}}^{\text{hourly}}$ are the hourly costs of floor space and repairs, respectively.

The CPU installation and acquisition cost is defined by

$$C_{a\&i}^{\text{CPU}} = N_{\text{GPPs}}(\text{pr}_{\text{GPP}} + T_{\text{GPP}}^{\text{ins}} S_{\text{tech}}) + \text{pr}_{\text{Sinf}}^{\text{CPU}}, \tag{5.19}$$

where pr_{GPP} is the price of the GPP and $T_{\text{GPP}}^{\text{ins}}$ is the installation time for the GPP.

The TRPs installation and acquisition cost is defined by

$$C_{a\&i}^{\text{TRP}} = N_{\text{TRP}} \left(\text{pr}_{\text{TRP}} + T_{\text{TRP}}^{\text{ins}} S_{\text{tech}} + \text{pr}_{\text{Fdrop}} \right), \tag{5.20}$$

where $T_{\text{TRP}}^{\text{ins}}$ is the TRP installation time, S_{tech} is the technician salary per hour, and pr_{Fdrop} is the price to install the final link from the Fiber to the Building (FTTB) infrastructure to the TRPs.

The fronthaul implementation cost can be dependent on various factors, like the type of the transmission medium, topology, number of derivation nodes, installed wired length, and distance between wireless nodes, among others (Yaghoubi et al., 2018; Farias et al., 2016; Monti et al., 2012). This chapter assumes that the fronthaul network utilizes a pre-deployed FTTB infrastructure, a reasonable assumption since the FTTB/FTTH penetration is already over 60 % in Europe and east Asia, growing more every year (Europe, 2023; Philpott; Fellenbaum; Frey, 2020). In this context, the only costs to deploy the fronthaul network are related to equipment at its tip, i.e., at the CPU and TRPs, being calculated as

$$C_{a\&i}^{\text{Xhaul}} = \sum_{l=1}^{N_{\text{TRP}}} \left(2\text{pr}_{\text{SFP}}^{F_{l,\text{peak}}} + \text{pr}_{\text{FEport}}^{F_{l,\text{peak}}} \right), \tag{5.21}$$

where $F_{l,\text{peak}}$ is the peak fronthaul bit rate for TRP l, calculated by $\sup_t F_{l,t}$. Moreover, $\text{pr}_{\text{Eport}}^{F_{l,\text{peak}}}$ is the price of the fronthaul Ethernet switch port capable of supporting rates of $F_{l,\text{peak}}$. Lastly, $\text{pr}_{\text{SFP}}^{F_{l,\text{peak}}}$ is the price of a grey SFP capable of supporting rates of $F_{l,\text{peak}}$ (Mahloo et al., 2014).

The total power consumption at time sample n is calculated through the power consumption at the associated time t by

$$P_t^{\text{total}} = P_{\text{TRP},t}\, M_{act,t} + P_{\text{CPU},t} + P_{\text{Xhaul},t}, \tag{5.22}$$

where $P_{\text{Xhaul},t}$ is the power associated with the backhaul/fronthaul network at the time t, which is calculated by

$$P_{\text{Xhaul},t} = \sum_{l=1}^{L_t} \left(2\text{pw}_{\text{SFP}}^{F_{l,\text{peak}}} + \text{pw}_{\text{FEport}}^{F_{l,\text{peak}}} \right), \tag{5.23}$$

where $\text{pw}_{\text{FEport}}^{F_{l,\text{peak}}}$ is the power consumption for an Ethernet fronthaul switch port capable of supporting rates of $F_{l,\text{peak}}$ and $\text{pw}_{\text{SFP}}^{F_{l,\text{peak}}}$ is the power consumption of a grey SFP capable of supporting rates of $F_{l,\text{peak}}$ (Fiorani et al., 2016).

The hourly repair costs are calculated by

$$C_{\text{rep}}^{\text{hourly}} = \sum_{i \in \mathcal{E}} \left(\frac{N_i\, N_{\text{tech}}^i\, (T_{\text{rep}}^i + 2T_{\text{trv}}) S_{\text{tech}} + \text{pr}_{\text{rep}}^i}{M_i} \right), \tag{5.24}$$

where $\mathcal{E}$ represents the set of different equipment types. This set is composed of the following elements: TRP, fiber final drop, SFP, GPP, rack networking device, fronthaul switch, and outdoor fibers. For a device of type i: N_i is the number of devices, N_{tech}^i is the number of technicians required for repair, T_{rep}^i is the repair time, pr_{rep}^i is the cost of replacement parts, M_i is the device's MTBF. Additionally, T_{trv} refers to the technicians' travel time (Mahloo et al., 2014).

The hourly floor space costs are calculated by

$$C_{\text{fSpace}}^{\text{hourly}} = \left(s_{\text{rack}} \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil + s_{\text{TRP}}\, N_{\text{TRP}} \right) \frac{\text{pr}_{\text{floor}}^{\text{year}}}{8760}, \tag{5.25}$$

where s_{TRP} and $\text{pr}_{\text{floor}}^{\text{year}}$ represent the physical area occupied by a TRP and the price of renting per year per unit of area, respectively. The number 8760 converts rent prices from yearly to hourly.

5.3 Numerical results

This section presents the results of this chapter. First, a reasonable baseline case study is defined and used to identify the main cost trends for distributed and centralized processing alternatives. Then, the impact of cost reduction in the non-CPU deployment infrastructure is evaluated, considering work-related expenditures. This evaluation aims to evaluate the benefits of markets with more affordable equipment and labor costs or by the adoption of the cheaper integrated solution cell-free mMIMO systems in the literature, like the one in (Interdonato et al., 2019). On the other hand, The prices of GPP and energy are also varied to identify possible changes in trends, as they can vary among vendors and globally, respectively. Finally, constructive parameters of the TRPs are varied to identify changes in cost trends, including the maximum number of UEs served by each TRP and its antenna count.

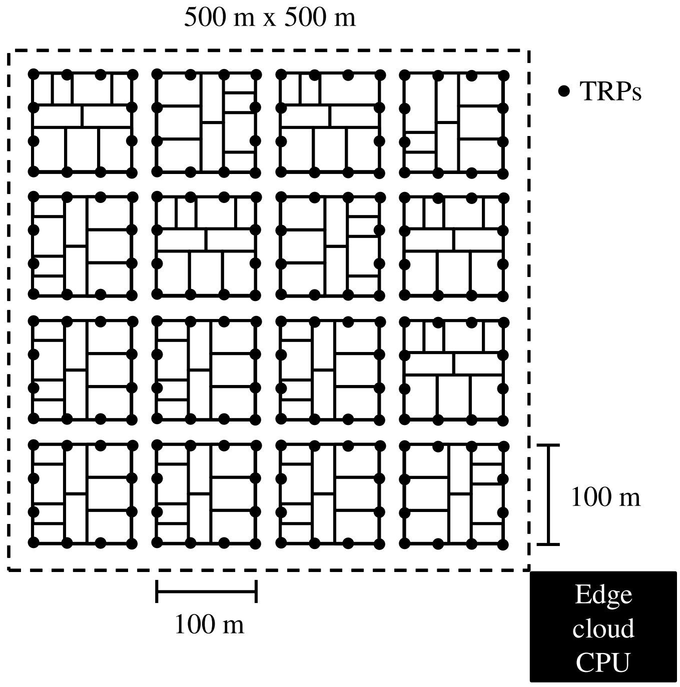
5.3.1 Case study

5.3.1.1 General assumptions

Figure 49 depicts the considered scenario, covering an area of 500 x 500 m with 16 blocks of buildings, each measuring 100 x 100 m. This scenario aims to emulate a dense urban environment. Although cities may differ in their building configurations, the grid building block is commonly found in larger cities like Barcelona or New York. Thus, it is considered a meaningful layout for a generic, dense urban environment. The TRPs are placed atop buildings at a 15 m height and are installed equally spaced between themselves on the side of each block. This configuration simplifies TRP deployment and is adequate to serve outdoor UEs on the streets, which are the focus of this chapter analysis. If, for any reason, the number of TRPs per block is not equal, some of them are randomly selected to have an additional TRP. Similarly, if the number of TRPs on each side of the block is unequal, one or more sides are randomly selected to have an additional TRP. Finally, UEs are randomly distributed on streets at 1.65 m height.

The number of active UEs fluctuates throughout the day according to a profile (Figure 50) with three possible levels of active UEs at different hours. Ideally, since the day is assumed to be discretized into hourly intervals, the profile should include 24 levels of active UEs. The main problem with this approach is that it is computationally burdensome since it would require 24 distinct simulations for each combination of precoders, UE demands, and TRP deployment strategies. Most simulations have a substantial count of TRPs and UEs and may take a long time to be executed, even in high-performance machines. Adopting only three possible levels is justified to depict a reasonable representation of active UE presence, capturing values at peak, valley, and approximate average while reducing the number of required simulations. Consequently, the adopted profile strikes a

Figure 49 – Considered urban-dense scenario. TRPs are placed atop buildings at a 15 m height.



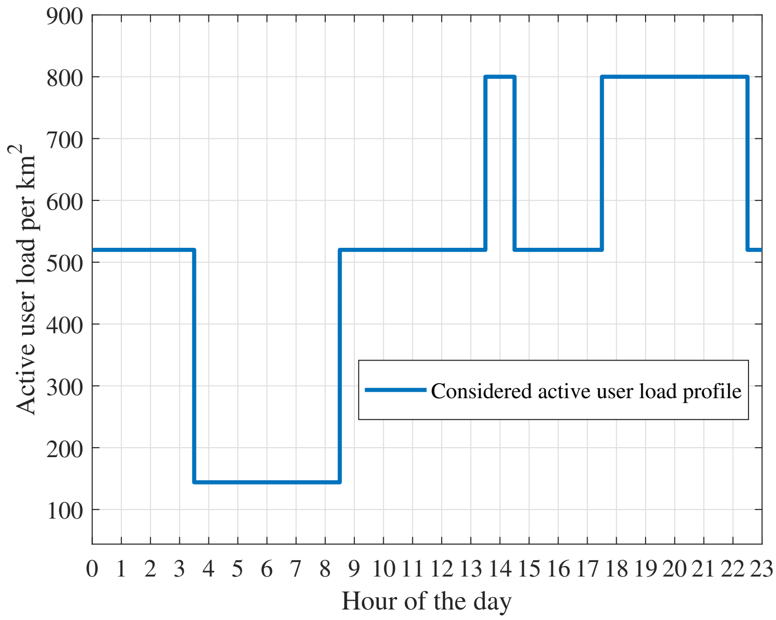
Source: elaborated by the author.

balance between the fidelity of portraying UE presence and the minimization of the computational resources required for simulation. The highest number of active connections occurs around 14:00 and 20:00, while the lowest number is around 6:00, resulting in a 5.6 peak-to-valley UE ratio. These figures align with the daily variation in the ratio of connected UEs to a LTE cell at a European metropolitan city (Trinh et al., 2017). The peak number of active UEs is calculated for a high-density urban area with 10,000 people per km^2, assuming each person has one UE. Furthermore, the calculation considers that the operator has a contract with approximately one-third of the UEs and that only outdoor UEs are served by the cell-free mMIMO network, which traditionally accounts for 25 % of all UEs (Ericsson, 2021).

Table 10 presents the power and price information for SFPs and Ethernet ports. The values are sourced from online network equipment suppliers. For Ethernet ports, values are extracted from the FS S8550, S8050, and S5850 switch families. For SFPs, Cisco devices with a 10 km range are used as benchmarks. All pricing is standardized using a CU equivalent to the cost of a grey optical 10 Gbps SFP, approximately US$27 at the time of the writing of this study. In this way, the prices for hourly technician salary (S_{tech}), kWh (pr_{kWh}), and yearly floor space rent ($\text{pr}_{\text{floor}}^{\text{year}}$) are specified as 7.4 CU, 3.7×10^{-3} CU, and 10.7 CU, respectively (Udalcovs et al., 2020).

Three processing strategies are compared: distributed LP-MMSE, centralized P-RZF, and centralized P-MMSE. The first follows a BTRP functional split, and the others follow

Figure 50 – Assumed profile of active UEs over the hours of the day.



Source: elaborated by the author.

Table 10 – SFP and fronthaul port price and power consumption.

Parameter/Equipment	**Price (CU)**			**Power (W)**		
Capacity (Gbps)	10	25	40	10	25	40
Grey SFP	1	2.6	11.4	1	1.3	3.5
Fronthaul Ethernet port	2.6	4.6	5.8	2.8	4.3	6

a BCPU functional split. All comparisons focus on the downlink performance, using the expected UE rate as the main parameter. Two distinct TRP deployment strategies are analyzed. The first deploys TRPs to achieve a given average UE rate and is not actively trying to provide fairness among UEs. While this strategy does not necessarily lead to unfair performance, it does not prioritize fairness. The second is based on an agreed-upon SLA rate and tries to emulate a deployment that actively tries to provide fairness. It deploys TRPs while ensuring that at least 40 % of the agreed rate is achieved at any time in 90 % of the coverage area.

5.3.1.2 System model assumption

The 3GPP Urban Micro (UMi) path-loss model is adopted for the system simulations (3GPP, 2020). The existence of LoS link components between every UE and TRP is checked by taking into account the positions of UEs and blocks of buildings in Figure 49. The LoS probability for the calculation of the Rician factor is given by the probability equations in (3GPP, 2020) for the UMi scenario. The correlation matrices follow the Gaussian local scattering model (Özdogan; Björnson; Larsson, 2019). A joint pilot assignment and TRP selection is assumed, where the first τ_p UEs are assigned mutually orthogonal pilots, and the remaining UEs are assigned to the pilot that experiences the lowest pilot contamination. Then, each TRP selects up to τ_p UEs with the highest average channel gain in each pilot (Björnson; Sanguinetti, 2020b).

Table 11 summarizes the system simulation parameters. Most are selected based on parameters commonly adopted in the literature (Demir; Björnson; Sanguinetti, 2021; Freitas et al., 2023; Klein, 2017; Dahlman; Parkvall; Skold, 2020). The number of antennas per TRP is chosen to represent the simplest TRP with multi-antenna processing capabilities. The assumed bit width of pilot samples and acceptable fronthaul data sample degradation assures a very low degradation in the channel estimates and data samples sent through the fronthaul. The maximum number of TRP connections per UE is selected to be high to ensure that each UE is connected to a large number of antennas. Lastly, the maximum number TRPs is chosen to allow an 8 m spacing between TRPs. This constraint is established to manage simulation computational requirements.

Table 11 – System, channel, and signal simulation parameters.

Parameter	**Values**
Number of TRPs at any SB l ($N_{l,m}$)	1
Number of antennas in any TRP m at SB l ($N_{l,m}$)	2
Number of supported UEs per TRP ($\max(\lvert\mathcal{D}_{l,m}\rvert)$) and uplink pilot samples (τ_p)	10
Coherence block samples (τ_c)	200
Carrier frequency	3.5 GHz
Bandwidth (B)	100 MHz
Number of subcarriers (N_{sc})	3300
Sampling frequency (f_s)	122.88 MHz
Symbol time (T_s)	35.38 μs
TRP total Tx power ($\mathrm{pw}_{\mathrm{Tx}}$)	23 dBm
UE total Tx power	20 dBm
Noise figure	7 dB
Angular standard deviation	15°
Shadow fading standard deviation	4 dB
Shadow fading decorrelation distance	9 m
Uniform linear array antenna spacing	half-wavelength
Fronthaul pilot samples bit width ($b^{\mathrm{pil}}_{l,m'}$)	10
Acceptable SE degradation due to fronthaul data samples (a_{deg})	0.1 bps/Hz
Maximum TRP connections per UE ($U_{\max}$)	64
Maximum deployed TRPs ($M_{tot,\max}$)	800

5.3.1.3 TRP model assumptions

The TRP's DSP power consumption and pricing are based on the TMS320C6671 /72/74/78 family by Texas Instruments. In this context, several key approximations have been outlined: the DSP core single precision processing capacity ($\mathrm{CAP}_{\mathrm{Dcore}}$) is 20 GOPS, the idle DSP core power slope ($\gamma_{\mathrm{pwDcore}}$) is 0.57 W/GOPS, the power slope related to processing load on the DSP (γ_{pwDSP}) is 49.1 mW/GOPS, the other related DSP power consumption ($\mathrm{pw}_{\mathrm{DSP}}^{\mathrm{other}}$) is 8.52 W, the price slope for DSP cores ($\gamma_{\mathrm{prDcore}}$) is 0.42 CU, and the base price of the DSP ($\mathrm{pr}_{\mathrm{DSP}}^{\mathrm{base}}$) is 2.92 CU.

A comprehensive breakdown of the pricing and power assumptions for various TRP subcomponents can be found in Table 12. Prices are based on an online electronic components supplier. Power consumption values are based on (Debaillie; Desset; Louagie, 2015). Besides that, the same price and power assumptions made for the fronthaul Ethernet switch ports are used for the I/O interface of the TRP.

Table 12 – Power and pricing assumptions for TRP components.

Component	**Price (CU)**	**Power (W)**
Filter	0.05	0.125
VGA	0.32	0.063
IQ modulator	0.39	0.2
DAC	0.14	0.175
ADC	0.14	0.225
Antenna	0.42	-

The price reduction factor due to SoC integration of the analog front-end (α_{SoC}) equals 0.44. This figure is derived from schematics of SoCs possessing similar subcomponents. The factor is calculated considering the pricing of these SoCs in an online electronic components supplier in relation to their discrete circuit counterparts.

Lastly, the price to install the final fiber drop from the building FTTB structure to the TRPs ($\mathrm{pr}_{\mathrm{Fdrop}}$) is 5.6 CU, which is based on the price of a drop in fiber internet installation for a building according to a telecommunication service company.

5.3.1.4 Edge clould CPU model assumptions

The deployed GPPs are based on Dell 1U PowerEdge R650xs rack servers, featuring a chipset with dual Intel Xeon Gold 6330 processors, one Solid-State Drive (SSD), and 16 sticks of 8 GB of Random Access Memory (RAM). This setup results in a power consumption of 242 W when idle ($\mathrm{pw}_{\mathrm{GPP}}^{\mathrm{idle}}$) and 652 W at peak ($\mathrm{pw}_{\mathrm{GPP}}^{\mathrm{peak}}$) operation (Principled Technologies, 2021). The Intel Xeon Gold 6330 has a base clock of 2 GHz, 28 cores, and an Ice Lake microarchitecture, supporting 64 single-precision FLOPS per cycle. The resulting GFLOPS capacity can be converted to GOPS by a factor of 1, resulting in a GPP with a GOPS capacity ($\mathrm{CAP}_{\mathrm{GPP}}$) of 7168 for a price ($\mathrm{pr}_{\mathrm{GPP}}$) of 367.7 CU.

For the racks, a 42U configuration is assumed. This means that, when utilizing a 1U server, the total capacity of each rack ($\mathrm{CAP}_{\mathrm{rack}}$) is 42 GPPs. Each rack requires a space (s_{rack}) of 1.728 m^2. From the pricing standpoint, the cost of acquisition and installation for both the rack and the accompanying network equipment ($\mathrm{pr}_{\mathrm{rk\&nt}}$) is 370.4 CU (Hardy et al., 2013).

For the support infrastructure to the IT components, the cooling PUE ($\mathrm{PUE}_{\mathrm{cool}}$) is 1.3, while the pricing for cooling and power distribution infrastructure ($\gamma_{\mathrm{Co|PD}}$) is 0.46 CU/W (Hardy et al., 2013; Cui et al., 2017). For the backup power solution, a battery bank is assumed. The acquisition and installation of each battery cost ($\mathrm{pr}_{\mathrm{bat}}$) is 11.11 CU, and their capacity ($\mathrm{CAP}_{\mathrm{bat}}$) is 1512 Wh (Jahid et al., 2020). The battery bank is designed to support an outage time of 5.52 hours, equal to the expected non-momentary energy interruption time in the United States. Finally, the inverter acquisition and installation price slope (γ_{inv}) is 0.015 CU/W (Jahid et al., 2020).

5.3.1.5 Installation and repair assumptions

The presented cost model requires TRPs and GPPs installation time. The first is assumed to be one hour. The second breaks down as follows: 30 minutes for physical server installation, 10 minutes for network connection, and 30 minutes for server provisioning, cumulatively amounting to 1.17 hours. These estimations are based on analogous components in other types of networks and the duration of manual server provisioning (Acatauassu et al., 2021; Chen et al., 2010; Principled Technologies, 2020).

Table 13 presents repair parameters for various equipment types. GPP MTBF and repair time metrics are sourced from server node failure data in large-scale computational clusters (Martino et al., 2014). Other values are derived from analogous components in different network types (Chen et al., 2010; Fernandes et al., 2022). Outdoor fiber MTBF scales with fiber length, which can be obtained as in (Fernandez; Stol, 2015) for a block scenario. The time to repair an SFP is considered equivalent to installing a port in a switch. Replacement parts' prices are assumed to be the same as acquisition prices. For GPP parts, costs are calculated by scaling component costs with respective failure rates and normalizing them with the GPP failure rate.

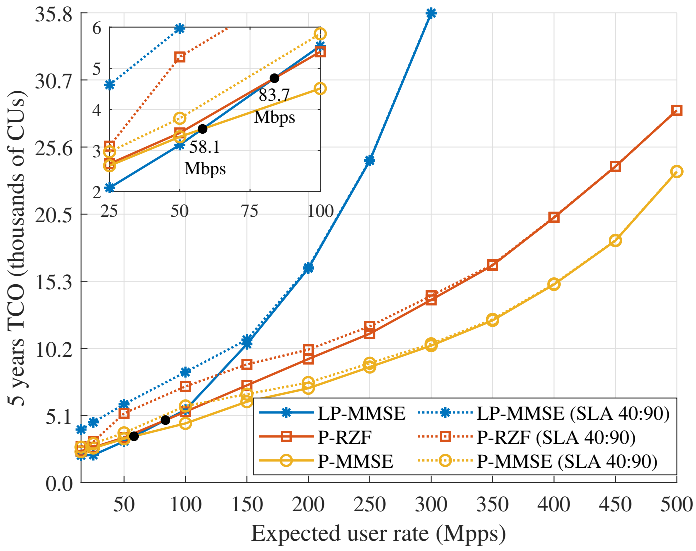
Table 13 – Installation and repair parameters.

Equipment	**Repair time (h)**	**MTBF (h)**	**Replacement parts price (CU)**
GPP	1.12	177523	100
SFP	0.17	2300000	$\mathrm{pr}_{\mathrm{SFP}}^{F_{l,\mathrm{peak}}}$
TRP	1	520000	$\mathrm{pr}_{\mathrm{TRP}}$
Fr. switch	7	500000	$\mathrm{pr}_{\mathrm{FEport}}^{F_{l,\mathrm{peak}}}$
Out. fiber	7	$1754386 \times$ km	–
Fiber drop	–	10000000	$\mathrm{pr}_{\mathrm{Fdrop}}$

For networking rack equipment, repair time, MTBF, and replacement parts cost are assumed to be equivalent to those of the fronthaul switch. The estimated travel duration for the repair team is one hour. Most repairs involve a single technician, but outdoor fibers require a trio (Chen et al., 2010).

5.3.2 Baseline results

Figure 51 – TCO after five years of operation concerning the expected UE rate for the case study assumptions ($N = 2$ and $\max(|\mathcal{D}_l|) = 10$). Intersection points between distributed and centralized processing under the same type of TRP deployment are marked by black dots. A zoom of the initial part of the curves is presented at the northwest part of the figure.



Source: elaborated by the author.

Figure 51 provides an overview of the TCO after five years of operation concerning the expected UE rate, which is calculated by summing (5.17) and (5.18). The analysis includes distributed LP-MMSE and centralized P-RZF and P-MMSE processing implementations under the case study assumptions, as outlined in Subsection 5.3.1. The cost differences between these implementations originate from the variations in the parameters within (5.17) and (5.18), which, in turn, are influenced by network requirements calculated in Section 5.1 for each type of processing. The data points span from expected rates of 15, 25, and 50 to 500 Mbps in increments of 50 Mbps, allowing for a detailed examination of the cost implications across a spectrum of UE demands. Additionally, the cost range is presented up to 35.8 thousand CUs, providing a comprehensive view of the economic considerations. Notably, the observed TCO trends exhibit exponential behavior concerning the expected UE rate, with distinct growth rates discernible among the various processing alternatives.

It is evident that LP-MMSE starts with lower costs but experiences a more accelerated cost growth rate than centralized alternatives. For instance, by increasing the expected rate from 50 Mbps[2] to 200 Mbps[3], the cost of LP-MMSE increases by up to 5.22 times. In contrast, a centralized P-MMSE implementation sees a cost increase of only 1.96 times between the aforementioned UE rates. This behavior suggests that centralized deployment can be more attractive and future-proof for next-generation networks[4]. The direct comparison between the processing alternatives reveals that a distributed LP-MMSE implementation is the most cost-effective alternative for UE demands up to 58.1 Mbps. Beyond that point, a centralized P-MMSE implementation becomes the least expensive. The centralized P-RZF implementation is always more costly than P-MMSE, regardless of the rate considered, being even less economical than LP-MMSE up to UE expected rates of 83.7 Mbps. Based on the results, it is more beneficial to use the distributed implementation approach for low demands per UE, i.e., required UE rates up to slightly over 50 Mbps. However, the centralized approach is more advantageous for medium and high traffic demands.

On the other hand, it is essential to note that these findings do not hold when considering a fairer Service Level Agreement where at least 40 % of the agreed UE rate is guaranteed to be achieved at anytime in 90 % of the coverage area (SLA 40:90) TRP deployment. In this case, the costs are always higher than in the previous analysis, and the curve behavior is initially increasing concave down before trending to the original exponential behavior in demands of 150 to 200 Mbps, even matching the non-SLA case starting in demands of 250 to 300 Mbps. In this way, LP-MMSE costs in lower demands are up to 104 % higher, while centralized processing alternatives have cost increases up to 36 %. In this way, for the SLA approach, centralized processing options are the most cost-effective for any expected rate, being the best way to implement a cell-free mMIMO system, with P-MMSE being the least costly processing alternative.

Figure 52 provides a comprehensive insight into the absolute value and cost composition of both CAPEX and OPEX across expected UE rates of 50, 100, 200, and 300 Mbps. These rates are achievable by all processing alternatives under the specified case study assumptions. Notably, results for the fairer SLA 40:90 TRP deployment are exclusively presented for P-RZF, as the behavior changes from the non-SLA results can be easily discerned by analyzing this specific precoder.

The findings underscore that CAPEX is the predominant factor in the five-year TCO for all expected UE rates and processing alternatives, representing between 73.2 % and 75.9 % of the costs. Extrapolating these results, it becomes apparent that for demands

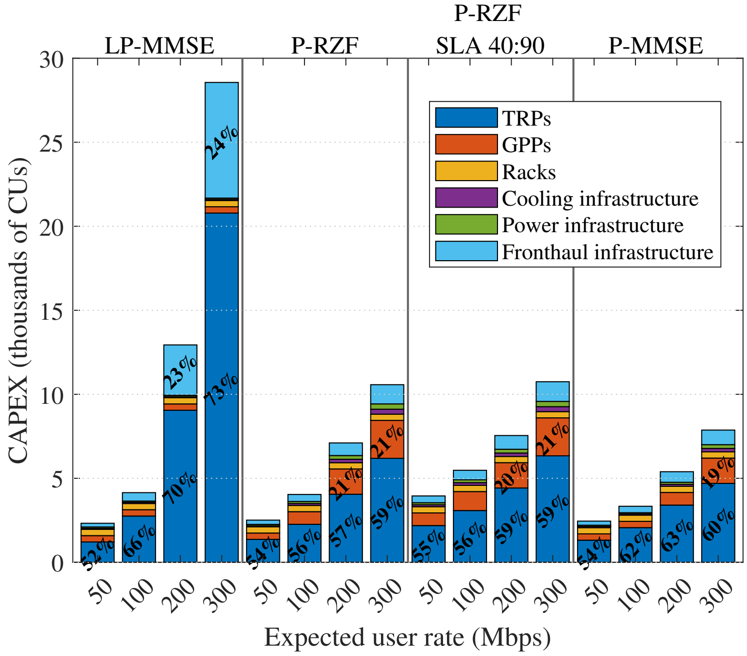
[2] The required 5G downlink UE rate for an urban wide-area scenario (3GPP, 2023). It can handle Full HD cloud Virtual Reality (VR) and 4K 3D video (HKT GSA; Huawei, 2019).

[3] A UE rate capable of handling most bandwidth-intensive applications, such as Augmented Reality (AR), cloud 2K VR, and 8K 3D video (HKT GSA; Huawei, 2019).

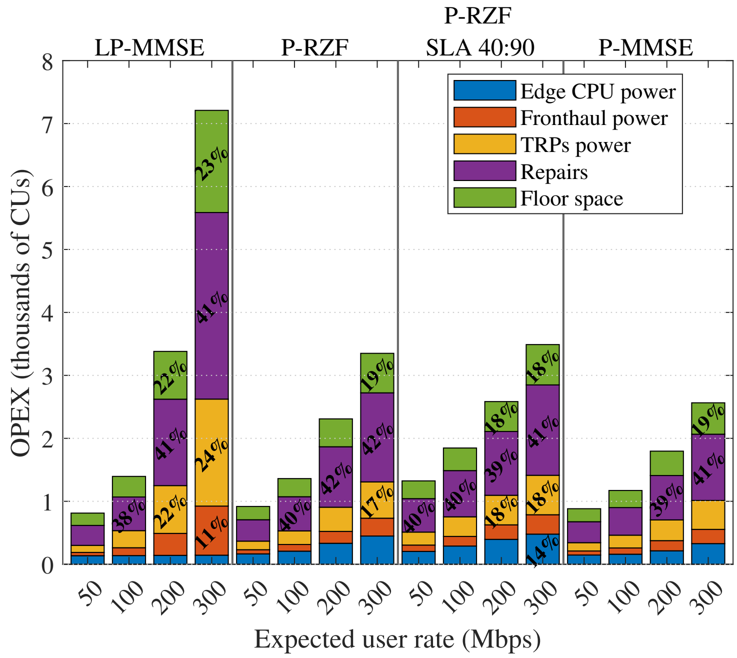
[4] In 6G systems, improvements should be sought as possible for uplink and downlink data rates within economic and sustainability constraints, since a 10x or 100x increase from 5G UE rates may be unsustainable (NGMN Alliance e.V., 2023).

Figure 52 – CAPEX and 5 years OPEX values and composition for up to six expected UE rates and the case study assumptions ($N = 2$ and $\max(|\mathcal{D}_l|) = 10$). The nominal non-SLA TRP deployment is considered unless when specified. The percentages within the stacks of bars represent the contribution of a component to the CAPEX or OPEX composition.

(a) CAPEX.



(b) 5 years OPEX.



Source: elaborated by the author.

of 50 Mbps, the total OPEX would reach the CAPEX value in 14.3, 13.7, and 13.9 years of operation for LP-MMSE, P-RZF, and P-MMSE, respectively. Furthermore, for demands of 300 Mbps, the total OPEX would equal the CAPEX value in 19.8, 15.8, and 15.3 years of operation for LP-MMSE, P-RZF, and P-MMSE, respectively. These results signify that

CAPEX remains the dominant factor in the TCO for an expected operation time ranging between 5 and 15 years, a typical duration for communication networks, especially in high-traffic demands scenarios.

Figure 52(a) illustrates the breakdown of CAPEX, highlighting its key components, including the acquisition and installation of: (i) TRPs, (ii) GPPs, (iii) GPP racks, (iv) cloud cooling infrastructures, (v) cloud power infrastructure, and (vi) fronthaul infrastructure. In the context of distributed LP-MMSE processing, the primary cost driver is related to TRPs, which presents a substantial increase in value and CAPEX participation with growing traffic demands. For instance, when the expected UE rate reaches 300 Mbps, the TRP cost alone accounts for 73 % of the total CAPEX. This growth can be attributed to the significantly larger number of TRPs needed to support higher UE rates effectively. The fronthaul cost becomes more relevant for LP-MMSE as the demands increase, with higher capacity transceivers in the fronthaul interface needed, constituting up to 24 % of the CAPEX at 300 Mbps. In contrast, the costs associated with cloud infrastructure for the distributed LP-MMSE implementation remain relatively minor, exhibiting no significant growth even with increased supported traffic demands.

Concerning the centralized processing implementations, P-RZF and P-MMSE share TRPs as the primary cost driver. Despite this, the dominance of TRP costs is less pronounced than in the distributed case, as it grows slower with supported traffic demands. Costs related to the GPPs also grow with increased supported traffic demands, going from negligible participation at 50 Mbps to around 20 % participation at 300 Mbps. It is noticeable that P-MMSE has lower costs than P-RZF due to reduced expenses in both TRPs and GPPs, originating from the higher performance of P-MMSE, which reduces the required number of deployed TRPs and consequently lowers processing complexity. Furthermore, it is worth noting that the expenses with fronthaul are comparatively smaller in centralized processing implementations than in the distributed one. This disparity is due to the fronthaul bit rate scaling with the number of antennas in the first case and UEs served by each TRP in the second (Björnson; Sanguinetti, 2020a).

When considering the fairer SLA 40:90 TRP deployment, it is noticeable that TRP and GPP costs are more elevated for all considered demands. In the cases of 50 and 100 Mbps, the cost increase compared to the non-SLA case is more pronounced. This fact is primarily attributed to the requirement for a higher number of TRPs to ensure fairness in lower demands, leading to increased processing computational complexity. As the demands approach 200 and 300 Mbps, the number of deployed TRPs in the non-SLA is sufficiently large to result in improved fairness, resulting in similar TRP and GPP costs to the SLA 40:90 case. This behavior explains why the SLA 40:90 TCO curve initially exhibits an increasing concave downtrend before trending towards the original exponential behavior of the non-SLA case.

Figure 52(b) provides a comprehensive breakdown of the yearly OPEX, highlighting its key components: (i) Edge CPUs power consumption, (ii) TRPs power consumption, (iii) fronthaul power consumption, (iv) repairs, and (v) floor space. Notably, the repair cost emerges as the largest contributor to the OPEX, accounting for between 38 % and 42 % of the total OPEX. It is followed by floor space and TRP power consumption, which can make up to 24 % and 23 % of the OPEX, respectively. Fronthaul power consumption is mostly negligible, except for LP-MMSE under higher demands. For instance, at 300 Mbps per UE, it reaches 11 % of the OPEX. The CPU power is mostly irrelevant for the distributed alternatives. In contrast, for the centralized ones, it becomes more relevant at medium-high rates, achieving up to 14 % of the OPEX in the 300 Mbps scenario.

The increase in most cost categories with UE demands is primarily driven by the growing number of deployed TRPs, leading to the increased deployed area, number of failures, computational complexity, and number of fronthaul connections. This behavior is also the reason why the fairer SLA 40:90 deployment incurs somewhat higher costs in all OPEX categories, especially in lower demands, since SLA 40:90 has more TRPs than its non-SLA counterpart. For higher demands, the behavior of the SLA and non-SLA deployments is mostly similar.

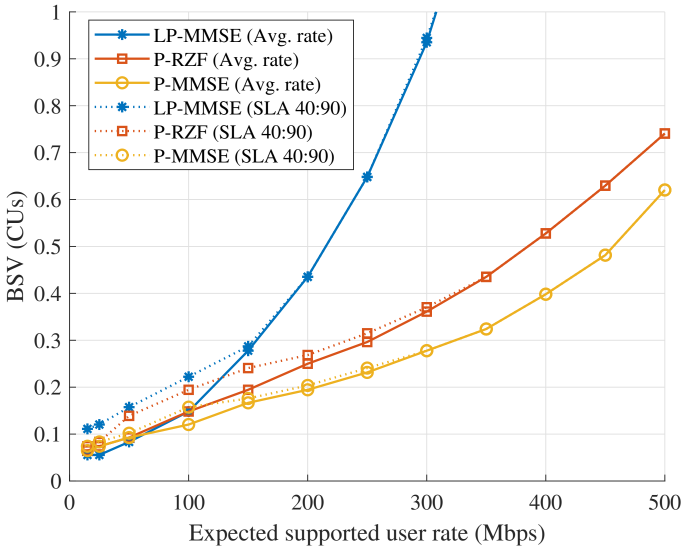
5.3.2.1 Evaluation of subscription prices to obtain profit

The analyses to this point have focused on TCO, CAPEX, and OPEX. These figures are ideal for comparative feasibility analysis of different deployment alternatives. Despite this, they do not fully reveal the feasibility in terms of profitability. A deployment option might appear more feasible than others in terms of TCO, but still be unfeasible in general due to lack of profitability, which can only be accessed considering metrics like RoI and NPV.

To address this, Figure 53 illustrates the monthly BSV required for the operator to achieve profits after five years of operation. The analysis includes distributed LP-MMSE and centralized P-RZF and P-MMSE processing implementations based on the assumptions outlined in the case study. Moreover, it is assumed that the number of subscribers equals the peak number of users in the area the operator may serve, including indoor users. The described case study in Section 5.3.1 system focuses only on outdoor service, but any indoor hotspot user can become an outdoor user at any time by exiting a building, being considered a potential user of the outdoor network. Then, based on the case study assumptions, the number of subscribers is equivalent to 800 users.

The BSV analysis provides an understanding of the actual feasibility of the compared solutions, as it provides the subscription prices for the system to be profitable. If these prices align with current subscription rates, it can be concluded that the deployment alternatives are financially feasible. For instance, in 2023, Western European prices per

Figure 53 – Monthly BSV for five years of operation concerning different expected UE rates under the case study assumptions ($N = 2$ and $\max(|\mathcal{D}_l|) = 10$).



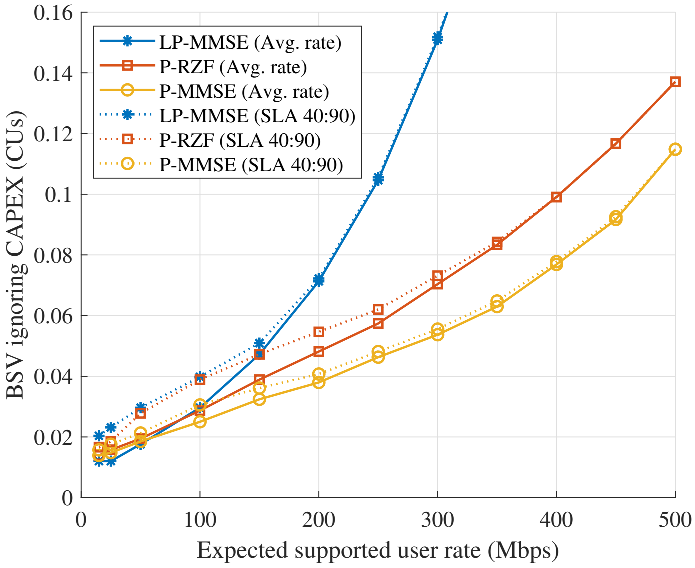
Source: elaborated by the author.

GB of mobile data averaged 0.077 CU(Cable.co.uk, 2024). Given the average consumption of 15 GB per user in the same region, the average plan price is around 1.15 CUs(Jakopin et al., 2023). Therefore, for the case study of this chapter, subscription prices should be close to this value to ensure feasibility. Preferably, the prices should be even lower to account for potential losses due to taxes and other financial expenses.

Consequently, Figure 53 is limited to a 1 CU subscription price. It is noticeable that the centralized option maintains subscription values within the proposed limit for any considered user demand, being feasible for all of them. As a highlight, P-MMSE requires a subscription of around 0.6 CUs for profit in 5 years for the highest considered demand, indicating that cell-free can be very competitive even under approximately half of the current average subscription fees in Western European countries. The distributed LP-MMSE can maintain a subscription fee below 1 CU when offering up to 300 Mbps to each user, which is enough for a truly enhanced mobile broadband experience in 5G networks (HKT GSA; Huawei, 2019). For higher user demands, the distributed option falls out of favor and can be considered unfeasible for profits under five years of operation.

If the CAPEX investment has already paid off, which is expected to happen within five years if the subscription fees from the previous results are adopted, the yearly BSV to ensure RoI becomes much smaller than the subscription fees from Figure 53. Figure 54 presents the results for this situation. It is clear that the distributed LP-MMSE can maintain fees as low as 0.16 CUs even with 300 Mbps per user, a fee significantly lower than the current fees practiced in developed countries. As for centralized processing, it ensures RoI even with fees under 0.14 CUs, even for the high-demand scenario of 500

Figure 54 – Monthly BSV for five years of operation, when ignoring CAPEX, concerning different expected UE rates under the case study assumptions ($N = 2$ and $\max(|\mathcal{D}_l|) = 10$). It is indicative of the required subscription monthly fee for yearly RoI.



Source: elaborated by the author.

Mbps per user at any time.

5.3.3 Impacts of price variations

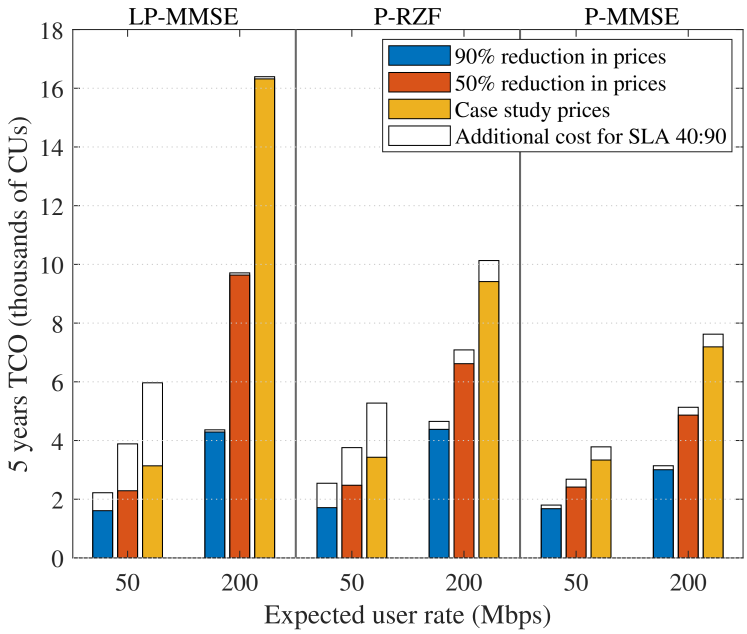
The prices of TRP, fronthaul infrastructure, GPP, and energy consumption play a crucial role in influencing both CAPEX and OPEX. Analyzing how variations in assumed case study prices affect overall costs is indispensable for making informed decisions regarding the cost assessment of a cell-free mMIMO network.

5.3.3.1 Non-CPU deployment price reduction

cell-free mMIMO systems stand to benefit from simpler and more affordable TRPs, especially in integrated solutions with low installation time and complexity, such as the one in (Interdonato et al., 2019). Additionally, some markets may benefit from these simpler TRPs even with non-integrated setups due to their manufacturing capabilities and lower labor costs. In both cases, the cost related to the TRPs might be more economical than the one obtained from case study assumptions. In other words, the considered market or an integrated solution has the potential to decrease all considered non-CPU acquisition and installation expenditures.

Figure 55 presents insights into the 5-year TCO for 90 %, and 50 % reductions in TRP and fronthaul prices, including work-related expenditures for network deployment, regarding the case study assumptions. These conditions aim to emulate the potential cost reductions from integrated cell-free mMIMO solutions that reduce installation time

Figure 55 – TCO for price variations in TRP and fronthaul prices concerning the case study, including equipment and work-related expenses. The aim is to emulate the potential cost reductions from integrated cell-free mMIMO solutions that reduce installation time and complexity, like the one in (Interdonato et al., 2019), or markets with cheaper labor and equipment. Colored bars represent costs for nominal non-SLA TRP deployment, while the colorless stacked bars depict the additional cost incurred by adopting the fairer SLA 40:90 TRP deployment.



Source: elaborated by the author.

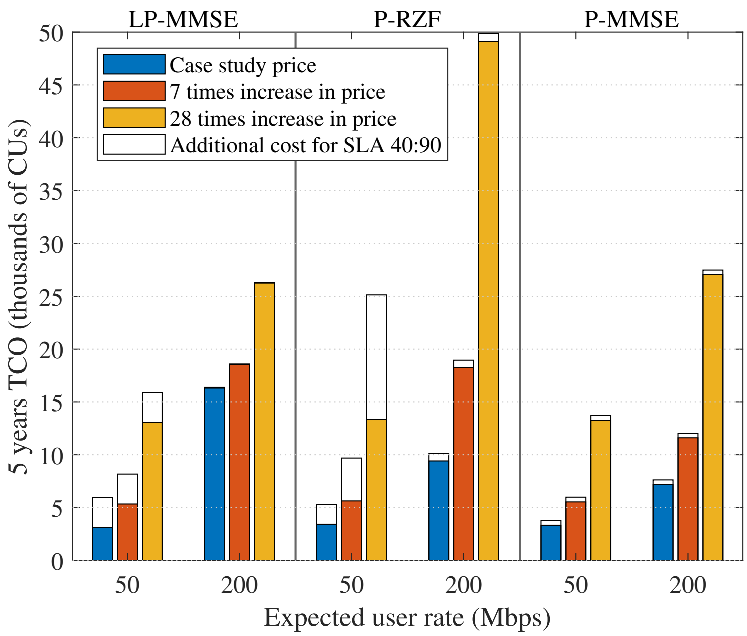
and complexity or markets with cheaper labor and equipment. Results representing low and medium demands are shown, equivalent to 50 and 200 Mbps per UE, respectively. From a purely economic perspective, the original findings remain the same despite price reductions. That is, LP-MMSE is the best approach in a non-SLA TRP deployment, and P-MMSE is the best choice in other cases. However, carefully examining the results reveals notable changes compared to the results of the case study prices. With an 85 % to 90 % reduction in non-CPU price variables, the distributed LP-MMSE becomes more economical than the centralized P-RZF in medium demands. Moreover, while P-MMSE remains the most affordable alternative in low-demand scenarios, it exhibits a very similar cost to LP-MMSE, hovering around 2 thousand CUs.

These results indicate that solutions or markets with reduced non-CPU equipment acquisition and installation costs, such as the integrated solution in (Interdonato et al., 2019), make distributed processing more cost-competitive if they provide an 85 % to 90 % reduction in non-CPU expenditures. Moreover, even if only a 50 % reduction is provided, such solutions or markets make the cell-free system significantly more affordable at higher rates, reducing costs in the demands around 200 Mbps per UE by multiple thousands of CUs, which are equivalent to 42 % to 75 % TCO reductions, depending on the processing scheme.

5.3.3.2 CPU deployment price reduction

Centralized processing implementations for cell-free mMIMO systems depend more on CPU component prices as the UE demands increase. The GPP prices assumed in the case study could be higher since the lowest price found in the conducted market research was considered.

Figure 56 – 5-years TCO for GPP price variations concerning the case study. Colored bars represent costs for nominal non-SLA TRP deployment, while the colorless stacked bars depict the additional cost incurred by adopting the fairer SLA 40:90 TRP deployment.



Source: elaborated by the author.

Figure 56 provides insights into the 5-year TCO for a seven and 28 times increase in GPP prices compared to the case study assumptions. Although these conditions surpass the identified range in the market research conducted for GPP prices, which had a maximum of 4 times increase, the analysis can offer valuable observations on the cost trends of different processing alternatives. The presented results represent low and medium demands, corresponding to 50 and 200 Mbps per UE, respectively. It is noticeable that an increase of seven times in GPP prices can elevate the TCO by 37 % to 83 % for low demands and 14 % to 93 % for medium demands. Notably, two significant changes were observed concerning the results of the case study findings. For both low and medium demands, LP-MMSE becomes more cost-effective or remains competitive relative to P-RZF, irrespective of the utilization of the fairer SLA 40:90 TRP deployment.

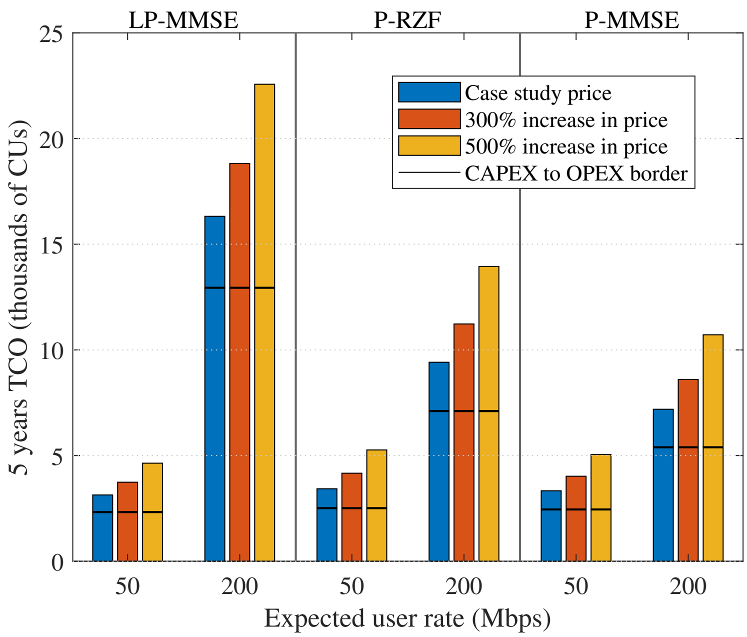
These results reaffirm the advantages of the more negligible dependence on CPU cost for distributed processing approaches. Concerning LP-MMSE, the cost increases for seven times GPP prices can be up to 1.5 and 9 thousand CUs higher for P-MMSE and P-RZF, respectively. Moreover, a further GPP price increase of 28 times can render the

LP-MMSE approach more affordable than the P-MMSE alternative in medium demands. Despite this, it is crucial to note that the occurrence of these changes in findings concerning the case study results depends on an CPU price increase of at least seven times. Thus, the market research increase of up to 4 times in prices cannot alter the findings from the case study results.

5.3.3.3 Energy price variation

Energy costs vary significantly based on deployment location. The case study employed a reference price for the kWh, which would be compatible with developed energy-rich countries where power is not so expensive. Despite this, developed European countries could have kWh prices up to 6 times higher at the date of this chapter submission. In this context, an analysis of the variation in energy prices is fundamental to ensure that the findings of this chapter can be applied to different economic realities.

Figure 57 – 5-years TCO for energy price variations concerning the case study. A line divides the participation of CAPEX and OPEX in the TCO. Only results for the nominal non-SLA TRP deployments are shown.



Source: elaborated by the author.

Figure 57 provides insights into the 5-year TCO for a 300 % and a 500 % increase in energy prices compared to the case study assumptions. Results representing low and medium demands are shown, equivalent to 50 and 200 Mbps per UE, respectively. A line is used to divide the participation of CAPEX and OPEX in the TCO. Values below the line account for CAPEX, and those above represent OPEX. For a more aesthetic presentation, results for the fairer SLA 40:90 TRP deployment are omitted, but the findings of non-SLA ones also apply to the fairer case. It can be observed that changing the energy price can significantly increase the TCO. A 500 % price increase can cause up to a 53 % increase in total costs. Despite this, there are no changes in the most and least cost-effective processing alternatives concerning the case study results.

The main change in relation to the case study results is the level of OPEX dominance on the TCO, which becomes much higher as the energy price increases. In fact, OPEX is almost the same as CAPEX for a 500 % price increase over five years of operation. In this situation, extrapolating the results shows that OPEX would reach the CAPEX value in 4.5 to 6.71 years of operation, depending on the processing alternative and demands. This makes OPEX the dominant factor in the TCO for the typical 5 to 15 years of operational life of communication networks. A more reserved but still significant increase in energy prices of just 300 % makes the OPEX reach the CAPEX value in 7.8 to 11 years of operation, depending on the processing alternative and demands, providing higher chances for OPEX dominance in the typical operational life. These findings justify works related to increasing energy efficiency in cell-free mMIMO systems.

5.3.4 Impact of UEs supported per TRP variation

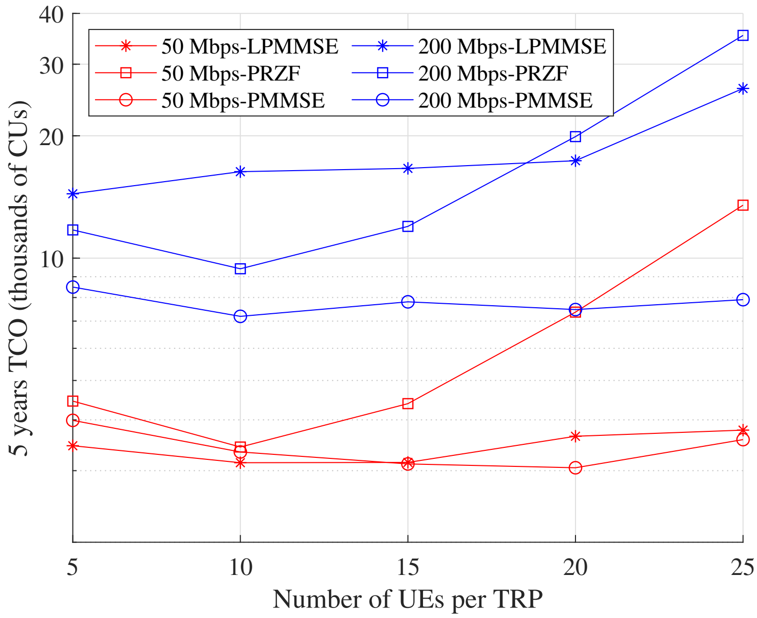
The number of supported UEs per TRP can strongly influence the performance and costs of different processing implementations. For example, the interference levels and computational complexity may experience substantial variations, especially for centralized processing. In this context, analyzing how variations in the number of supported UEs per TRP impact the TCO is essential to make informed decisions regarding processing implementations.

Figure 58 provides an overview of the TCO after five years of operation concerning the number of supported UEs per TRP for the expected UE rates of 50 and 200 Mbps, representing low and medium demands. Results for 5, 10, 15, 20, and 25 UEs per TRP are shown in two subplots representing (a) nominal non-SLA and (b) fairer SLA 40:90 TRP deployments. Moreover, besides the UEs supported per TRP variation, all other parameters are the same as in the case study. It can be noticed that from 15 UEs per TRP onward, SLA and non-SLA costs are almost the same for the centralized P-MMSE and P-RZF alternatives. For the distributed approach LP-MMSE, the cost difference between TRP deployments is significant in low demands but very similar in medium demands.

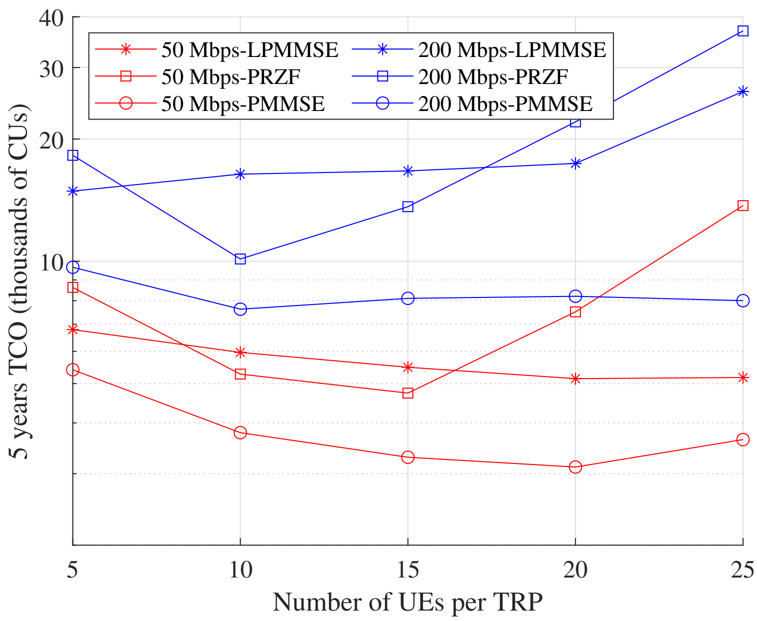
Figure 58(a) provides a detailed overview of the nominal non-SLA results. Notably, for low demands, the distributed LP-MMSE emerges as the most competitive implementation for up to 15 UEs per TRP. Beyond this point, P-MMSE becomes the preferred alternative. In the case of medium demands, P-MMSE consistently outperforms other alternatives by a substantial margin. An interesting behavior is the presence of a valley in the P-RZF curve, occurring at 10 UEs per TRP for both low and medium demands within the considered values of UEs per TRP. These findings suggest that the optimal

Figure 58 – 5-years TCO vs. maximum UEs per TRP for 50 and 200 Mbps expected UE rates, representing low and medium demands, respectively. Five values of maximum UEs per TRP are considered: 5, 10, 15, 20, and 25. Other parameters remain the same as in the case study.

(a) Nominal non-SLA TRP deployment.



(b) Fairer SLA 40:90 TRP deployment.



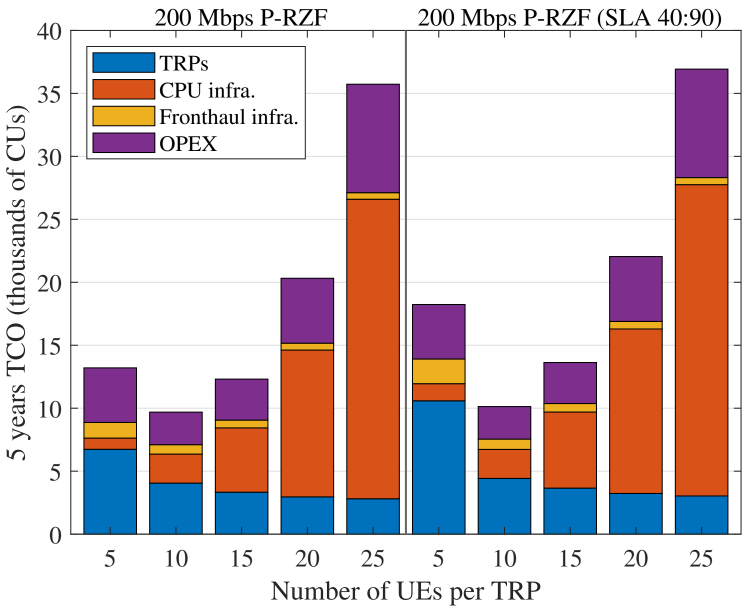
Source: elaborated by the author.

operation in terms of cost for P-RZF lies around 10 UEs per TRP. Moreover, it is shown that the concave-up behavior of the P-RZF can make it more expensive than LP-MMSE in medium demands, as seen in 20 and 25 UEs per TRP. The other processing alternatives exhibit a more uniform behavior, with minor variations attributed to changes in deployed

TRPs, computational complexity, and fronthaul requirements. Notably, the most significant variation outside of P-RZF occurs in the 200 Mbps LP-MMSE between 20 and 25 UEs per TRP. This variation is primarily due to fronthaul requirements scaling with the number of UEs per UE in distributed processing implementations.

Figure 58(b) provides detailed results for fairer SLA 40:90 TRP deployments. Notably, for both considered demands, P-MMSE emerges as the most competitive implementation regardless of the number of UEs served per TRP. The P-RZF curve exhibits a valley, as observed in the non-SLA results, occurring at 15 and 10 UEs per TRP for low and medium demands, respectively. Comparing it to the non-SLA results, there is a shift in the valley's location from 10 and 15 UEs per TRP. However, the cost difference between these points is small enough to say that for low demands, the optimal point of operation lies within this range. Additionally, another noteworthy change concerning non-SLA results is that P-RZF becomes more expensive for low UE counts per TRP being more economical than LP-MMSE only for 10 and 15 UEs per TRP. This behavior is attributed to the higher costs associated with the SLA 40:90 deployment, coupled with the concave-up nature of the P-RZF curve.

Figure 59 – TCO composition of P-RZF for demands of 200 Mbps per UE concerning 5, 10, 15, 20, and 25 maximum UEs per TRPs. Other parameters are the same as the case study assumptions. CAPEX is further divided into TRPs, CPU, and fronthaul costs.



Source: elaborated by the author.

Figure 59 presents the cost composition of the TCO concerning TRPs, edge CPU, fronthaul infrastructure, and OPEX for the 200 Mbps P-RZF curve in the supported UEs per TRP variation analysis. The aim is to better understand the concave-up behavior of the cost curve. It can be observed that for both Nominal non-SLA and fairer SLA 40:90

TRP deployments, the cost with TRP decreases. This reduction occurs because fewer TRPs are needed to support UE demands as UEs per TRP increase. However, the costs associated with edge CPU experience a significant increase with UEs per TRP. This is attributed to the higher number of common UEs between TRPs, leading to an increase in the computational complexity of partially centralized precoders/combiners, such as P-RZF and P-MMSE. While the valley phenomenon is evident for P-RZF, a similar trend is expected for P-MMSE. However, larger variations in UEs per TRP need to be observed to determine the point at which this occurs conclusively. The presented analysis of up to 25 UEs per TRP revealed minor variations, but it was inconclusive regarding the valley's location.

5.3.5 Impact of antennas per TRP variation

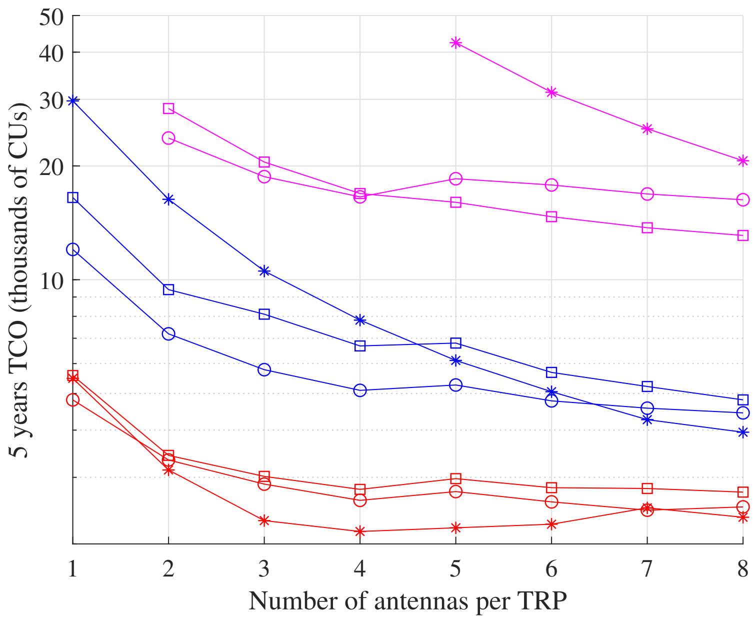
The number of antennas per TRP can strongly influence the performance and costs of different processing implementations. For example, distributed processing techniques are known to combat interference much better if the TRPs have more antennas. In this context, analyzing how variations in the number of antennas impact total costs is essential to make informed decisions regarding processing implementations.

Figure 60 provides a comprehensive overview of the TCO over a five-year operational period, considering different numbers of antennas per TRP for expected UE rates of 50, 200 Mbps, and 500 Mbps, representing low, medium, and high demands, respectively. The results for 1 to 8 antennas per TRP are presented in two subplots, depicting (a) nominal non-SLA and (b) fairer SLA 40:90 TRP deployments. All other parameters remain consistent with the case study. For high demands, the curves start in 2 antennas for centralized P-RZF and P-MMSE, and 5 antennas for distributed LP-MMSE. These are the minimum number of antennas where it becomes feasible to support 500 Mbps UE demands under the assumptions of the case study, considering the different processing schemes. Finally, it is important to note that this is the first result demonstrating the capability of distributed LP-MMSE processing to support demands of around 500 Mbps per UE.

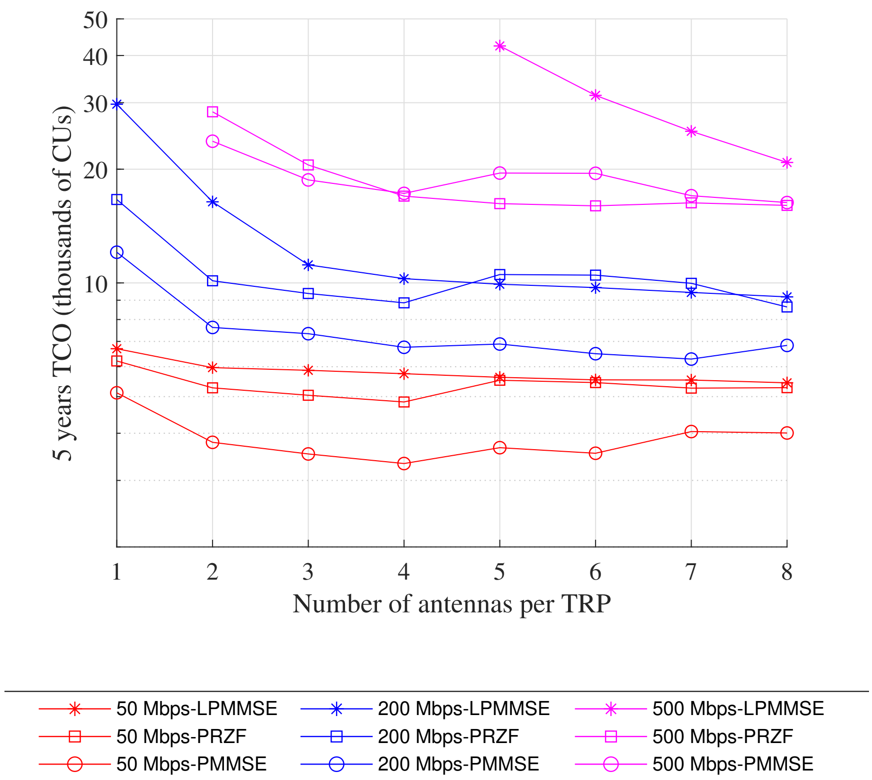
A comparison between non-SLA and fairer SLA 40:90 TRP deployments reveals that as the number of antennas increases, the latter becomes progressively more expensive than the former. This trend is primarily attributed to a more significant reduction in deployed TRPs in the non-SLA case with increasing antennas per TRP. Thus, when fairness is not explicitly addressed, providing higher rates with far fewer TRPs becomes possible, as the total number of antennas in all TRPs tends to remain similar. However, this behavior results in less evenly distributed TRPs across the coverage area, reducing macrodiversity and fairness. This explains why far more TRPs may be needed for fairer SLA 40:90 TRP deployments when considering a higher antenna count. The only exception

Figure 60 – 5-years TCO vs. antennas per TRP for 50, 200, and 500 Mbps expected UE rates, representing low, medium, and high demands, respectively. Other parameters remain the same as in the case study.

(a) Nominal non-SLA TRP deployment.



(b) Fairer SLA 40:90 TRP deployment.



Source: elaborated by the author.

to this behavior is LP-MMSE under 500 Mbps demands, which already has a TRP count high enough to provide fairness.

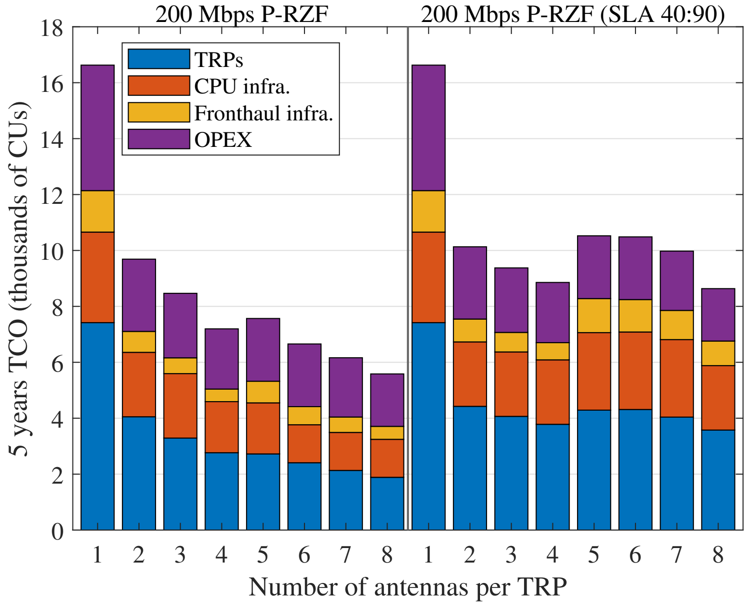
Figure 60(a) provides a detailed overview of the nominal non-SLA TRP deployment results. Notably, for low demands, the distributed LP-MMSE emerges as the most competitive implementation, starting from 2 antennas per TRP and having similar costs to P-MMSE at 7 antennas per TRP. Centralized P-MMSE is the most affordable for medium demands until 6 antennas per TRP. Beyond this, LP-MMSE becomes the most cost-efficient alternative. For high demands, P-MMSE is the more economical approach to up to 4 antennas per TRP. After this point, using P-RZF is more cost-effective. Focusing on distributed LP-MMSE, it can support 500 Mbps demands but is generally more expensive than the centralized approach, being 7.5 thousand CU more expensive with 8 antennas specifically. As for the centralized approaches, they mostly exhibit an interesting behavior from 4 to 5 antennas, where the cost increases instead of decreasing. Consequently, for high demands and P-MMSE, a minimum of 8 antennas per TRP is necessary to obtain a TCO smaller than in the 4 antenna case, despite the cost decreasing since 5 antennas per TRP.

Figure 60(b) provides detailed results for fairer SLA 40:90 TRP deployments. Notably, centralized P-MMSE proves to be the most cost-effective approach for low and medium demands across all considered numbers of antennas. For high demands, P-MMSE starts as the more economical option but loses its cost advantage to P-RZF after 4 antennas per TRP. Although it becomes close again, starting from 7 antennas, it never becomes less expensive than P-RZF. There are interesting behaviors for centralized P-MMSE, and P-RZF observed once again, particularly the transition from 4 to 5 antennas, which appears to increase costs in most of the analyzed conditions. Similar behaviors also occur at low demands for P-MMSE from 6 to 7 antennas, at medium demands for P-MMSE from 7 to 8 antennas, and at high demands for P-RZF from 6 to 7 antennas. These behaviors ensure that for centralized P-MMSE and P-RZF, costs with 4 and 8 antennas are similar for medium and high demands. Moreover, 4 antennas per TRP is the point where the lowest cost of the low demands is achieved by P-MMSE.

Figure 61 presents the cost composition of the TCO concerning TRPs, edge CPU, fronthaul infrastructure, and OPEX for the 200 Mbps P-RZF curve in the number of antennas per TRP variation analysis. The aim is to better understand the increasing behavior that sometimes occurs between two antenna counts in the cost curves, most often in the transition from 4 to 5 antennas per TRP.

In the nominal non-SLA case, it is observed that the cost with TRPs remains roughly the same when transitioning from 4 to 5 antennas per TRP. This implies that despite the reduction in the number of deployed TRPs, the price of an individual TRP increased significantly, offsetting any potential economic gains. The individual cost of a TRP always rises with the number of antennas, as more expensive analog front-ends and digital signal processors are needed to support higher antenna counts. In the case

Figure 61 – TCO composition of P-RZF for demands of 200 Mbps per UE concerning a variation of one to eight antennas per TRP. Other parameters are the same as the case study assumptions. CAPEX is further divided into TRPs, CPU, and fronthaul costs.



Source: elaborated by the author.

of centralized processing, the I/O interface of TRPs can also become more expensive as the fronthaul bit rate scales with the number of antennas. Moreover, for the same reason, fronthaul costs can increase. Thus, in the transition from 4 to 5 antennas per TRP, the fronthaul costs increased because the reduction in deployed fronthaul infrastructure from having fewer TRPs is insufficient to compensate for the increase in costs from the individual fronthaul equipment needed to support a higher bit rate. This transition is not observed for every antenna count because the capacity boundary between the considered transceivers is high. For example, a 14 or 24 Gbps fronthaul demand requires a 25 Gbps transceiver, but as soon as the fronthaul demand surpasses 25 Gbps, 40 Gbps transceivers, which are more expensive, need to be used.

For the fairer SLA 40:90, the explanation for the intermediate increases is similar, but an increase in CPU costs is also observed. This behavior happens because the reduction in TRP count is insufficient to compensate for the increased computational complexity introduced by the higher antenna count, as noticed in the transition from 4 to 5 antennas. Increasing the number of antennas should cause more computational complexity in centralized precoders' calculations. The decrease in CPU costs observed in most antenna count transitions occurs because the global computational complexity of the precoders decreases with fewer TRPs deployed.

5.4 Chapter summary

This chapter introduced a comprehensive cost assessment methodology to calculate the TCO of cell-free mMIMO networks. In this context, the study presented in this chapter contemplates the second proposal and hypothesis of the thesis, detailed in Subsection 1.3.2. The methodology includes models for network deployment, computational baseband processing requirements, fronthaul signaling, equipment pricing, and power consumption.

The network deployment model was based on a proposed TRP distribution method bounded by coverage or capacity constraints. The latter provided a user rate derived from the network average UE rate or a proportional fairness-based UE rate complying with a SLA. This approach ensures a fairer TRP deployment, maintaining a significant portion of this UE rate throughout a large part of the coverage area.

Together with the cost assessment methodology, price and energy consumption models for Edge CPUs and TRPs were proposed. These models allow the determination of the acquisition cost of the equipment in a generic, non-vendor-specific way. Additionally, they enable the computation of energy consumption under different processing and fronthaul loads at the equipment. Finally, a complete cost model was proposed, considering the acquisition and installation of equipment and links, maintenance, floor space rent, and power consumption.

The case study carried out in this chapter focuses on comparing distributed and centralized processing functional split options in a dense urban deployment in a developed country. The considered processing approaches were the centralized P-MMSE and P-RZF precoders, as well as the distributed LP-MMSE precoder. The results demonstrated that centralized processing implementation was generally more cost-effective, especially for an actively fairer TRP deployment. If fairness is not a concern, distributed LP-MMSE was the most feasible option for low user demands, around 50 Mbps per user. Additionally, a higher TRP antenna count, such as eight or more, or an 85 to 90% cost reduction in the CAPEX related to TRP deployment, was able to make the distributed processing implementation more cost-effective for medium demands, considered to be around 200 Mbps.

The results for the two signal processing solutions considered for centralized implementation (P-MMSE and P-RZF) showed that P-MMSE was definitely more cost-effective in low and medium demands. On the other hand, P-RZF could be more cost-effective in high-demand scenarios, with 500 Mbps per user, when the antenna count per TRP is higher than 4. However, this is contingent upon the number of UEs per TRP since it is shown that P-RZF achieved its minimum cost when each TRP served around ten UEs. Then, in a general sense, P-MMSE appears to be the more feasible solution.

A BSV analysis evaluated the required user subscription fee for a cell-free mMIMO network to be profitable in five years. It was shown that with centralized processing, the required subscription fee is below the current practiced values for mobile networks, even for high demands. In fact, the fees can be as low as half of the prices practiced in developed European countries. On the other hand, distributed processing was shown to be profitable within five years, under current practiced values, for demands up to 300 Mbps. Finally, after CAPEX is paid off, it was demonstrated that much smaller fees could be practiced to ensure RoI.

A sensitivity analysis of the prices and expenditures assumed in the case study was performed. It was revealed that implementations with significantly reduced fronthaul and TRP deployment costs provided remarkable TCO reductions for all processing alternatives. Moreover, These deployment cost reductions also had the potential to make distributed solutions more cost-competitive for medium demands. Further sensitivity analyses indicated that substantially higher-than-normal GPP prices are required to make centralized implementations less competitive. Finally, high energy prices do not appear to change the cost competitiveness level of the processing alternatives but could strongly reduce CAPEX dominance in the TCO.

Considering all findings, the centralized implementation utilizing P-MMSE precoding stands out as the most economically viable solution for cell-free mMIMO networks. This approach offers a reasonable cost across various user rates, allowing an affordable user subscription fee to ensure profits over five years, and making the network more future-proof than the distributed alternative. Besides that, thanks to its superior interference cancellation capabilities, P-MMSE maintains its cost-effectiveness even with simpler TRPs that have fewer antennas. While certain conditions may make distributed processing or P-RZF precoding more economically feasible, centralized P-MMSE generally offers superior economic benefits.

6 Conclusions and direction for future research

This thesis explored the deployment feasibility of cell-free mMIMO networks. It presented two innovative proposals, each featuring a distinct analytical tool to assess this feasibility. The first tool was a framework designed to evaluate network reliability concerning hardware failures in fronthaul connections and radio equipment, highlighting their impact on user rate performance. The second tool was a techno-economic framework designed to evaluate different types of deployments for cell-free mMIMO networks.

Both proposed frameworks offer robust methodologies for identifying feasible deployment strategies in scenarios of interest to society and industry, such as densely populated urban areas and office environments. Contributing to the practical application and efficiency of cell-free mMIMO networks as mobile communication systems.

The reliability framework modeled the network as a graph to instantaneously evaluate the consequences of hardware failures in fronthaul segments or equipment on interconnected devices, while considering the existence of redundancy equipment ensuring an alternative connection route in the case of failures. Additionally, the framework used a CTMC, constructed using the failure rates and repair times of the equipment and fronthaul segments, to represent the effects of cumulative failures. The CTMC was solved in a cumulative or up-to-absorption way using MCMC techniques. The proposed simulation strategy aims to assess degradation in SE over a set amount of time in the first case, and the time required to achieve a given SE degradation in the second case.

The application of the reliability framework to an indoor office scenario with both contaminated and uncontaminated pilots demonstrated that the impacts of individual failures could have a poor reliability level, justifying the deployment of protection schemes to offset the effects of failures in cell-free mMIMO with segmented fronthaul. It was shown that with half of the TRPs being serially connected, large SE performance reductions were obtained under single failure scenarios (around 20 to 30% on average). Moreover, It was noted that failures in the initial 40% length of the SB always cause a reduction equal to or larger than 20% in SE. On the other hand, when considering one-fifth of the TRPs serially connected, the impacts of a single failure are reduced but still could cause a reduction of 10% in SE. Besides that, the cumulative impacts of continuous failures over time still significantly affected the system SE, again justifying protection schemes.

Protection solutions to offset the effects of failures were developed for cell-free mMIMO. A wireless or wired interconnection was proposed for the integrated system known asERSS. For non-integrated systems, three fronthaul protection architectures were proposed: FD, PD, and CC, which were compared with an NP deployment. All solutions mitigated the impacts of any single failure on the system. Regarding cumulative failures, the ERSS protection solution extended the time until a 20% SE degradation without maintenance or supervision by 3 to 4 times. The non-integrated solutions increased this time by 1.5 to 4 times. Another way to interpret these results is that the time to schedule maintenance can increase by this scale, meaning it would take between 1 and 17.5 more years to require maintenance. This latter value may appear too large, but it was obtained for an FD solution with lots of redundancy equipment.

In the non-integrated cell-free mMIMO context, it was determined that CC was the most advantageous option when four or more SBs were available. Before this point, PD was more feasible. As for FD, it provides much higher reliability, but even under a simple non-compressive cost analysis, it is clear that its benefits do not justify its expenses. PD and CC offer a much better balance between additional redundancy equipment and provided reliability.

The techno-economic framework proposed a cost assessment methodology to calculate the TCO for cell-free mMIMO networks deployments across various user demands, taking into account fronthaul bandwidth limitations, TRP, and CPU processing requirements and capacities, considering distributed and centralized processing architectures. The framework could analyze the network for any scenario and consider variations at the active user connection load, being capable of representing variations on fronthaul and processing requirements during a normal operational day.

A complete cost model was proposed to be used with the methodology, covering both CAPEX and OPEX considerations. For CAPEX, expenses considered the acquisition and installation of (i) TRPs, (ii) edge cloud CPU, and (iii) fronthaul equipment. On the OPEX side, expenses took into account (i) repairs, (ii) equipment occupied floor space rent, and (iii) power consumption. Besides that, technician salaries impact both CAPEX and OPEX.

Equipment distribution and resource allocation models for cell-free mMIMO networks were proposed. They effectively allowed adequate network dimensioning, and evaluations of fronthaul limitations, considering available bandwidth and quantization aspects. Besides that, the models also clarified TRP and CPU processing capacities.

To achieve adequate network dimensioning, the model proposed for TRP considered user demands based on supported average user rates or designed towards fairness to ensure that a reasonable user rate was guaranteed across a large portion of the coverage area. When evaluating fronthaul requirements, a proposed model calculated the number of bits representing samples transmitted through the fronthaul to minimize degradation in SE compared to the non-constrained fronthaul case. In this context, the fronthaul requirements were modeled to ensure minimal losses in user rate due to fronthaul quantization. Furthermore, for the computational hardware requirements at the CPU and TRP, a model is proposed for two functional splits found in the cell-free mMIMO literature: BTRP and BCPU. The first favors distributed processing, with channel estimation and precoding/combining performed at the TRP, and the other favors centralized processing, with these functions performed at the CPU.

The application of the cost assessment methodology in a dense urban scenario indicated that centralized processing was the most economically feasible solution for cell-free mMIMO networks. Distributed approaches in scenarios where fairness was not a concern were less costly for lower demands (50 Mbps per user). Even then, it was shown that centralized options become progressively more cost-effective than distributed ones as user data rates increase, making centralized processing more future-proof. A BSV analysis showed that distributed option deployments could be profitable in five years under user subscription fees similar to current ones for demands up to 300 Mbps per user. Even then, the centralized option can do the same with much higher demands. For instance, in 500 Mbps demands per user, approximately half of the current adopted subscription fee would be enough to ensure profits.

Sensitivity analyses revealed that a strongly reduced expenditure in non-CPU related equipment (85% to 90%) and antenna count on TRPs (6 or more) could make distributed processing the more feasible option when user demand increased to 200 Mbps. The mentioned strong cost reductions are more likely in countries with lower labor costs and substantial manufacturing capabilities or in integrated cell-free mMIMO systems like the ERSS. Even then, for high demands, around 500 Mbps per user, centralized processing remained supreme as the most cost-effective approach.

Among the precoder/combiner options for centralized processing, P-MMSE was generally the most economical. Despite this, P-RZF was more cost-effective in higher-demand scenarios for TRPs with more than four antennas and a specific user count per TRP around 10. Counts higher or lower resulted in increased costs for P-RZF, which can be even worse than the costs of distributed approaches.

In summary, this thesis presented and validated two distinct analytical tools designed to evaluate the deployment feasibility of cell-free mMIMO networks. The performed analyses on network reliability and cost-effectiveness provided key insights into the strategic planning of cell-free mMIMO networks, identifying deployment alternatives that enhance the feasibility of these next-generation systems. The insights and tools, along with some models developed for them, are novel contributions to the field, making the work conducted in this thesis a benchmark for future research. Building upon the comprehensive analysis conducted in this thesis, the directions for future research include:

- Analysis of the reliability tool with more advanced precoders such as LP-MMSE, P-MMSE, and P-RZF, including hybrid centralized-distributed processing alternatives that can operate only in serially connected systems.
- A more in-depth evaluation of cross-connection-based protection schemes, considering the impacts on spectral efficiency that fronthaul limitations in the connections can cause.
- Considering different access mediums and topology options in both cost and reliability analyses of non-integrated cell-free mMIMO systems, especially since this thesis primarily focused on fiber solutions in the mentioned situation.
- Analysis of costs with multiple edge CPUs, including the dynamics of backhaul links interconnecting them.
- Approximation of the performed analyses to industry standards by evaluating a system operating under the functional splits proposed by 3GPP or the O-RAN standard. These functional splits have similarities with the literature on cell-free splits and may be adapted to serve this type of network.
- Comparison of the costs of a cell-free mMIMO setup with traditional cellular setups using both distributed and centralized signal processing solutions like single-cell and multi-cell minimum mean square error combining.
- Further investigation into the fronthaul limitation model used in the cost assessment methodology. For instance, the model can be improved by individualizing the bit width for data samples in different TRPs.
- Exploration of a paradigm shift where the central processing unit is envisioned as a cloud service hosted within a third-party data center, implying an edge CPU that the operator does not own.
- A merger of economic and reliability studies to evaluate with precision the cost of the protection alternatives, while a simple cost figure was presented in Chapter 4, it would be beneficial to adopt the comprehensive model of Chapter 5 to obtain more precise results.

Bibliography

3GPP. *Study on CU-DU lower layer split for NR, Release 15*. 2012. V15.0.0. Cited on page 169.

3GPP. *Study on channel model for frequencies from 0.5 to 100 GHz*. 2020. TR38.901, Version 16.1.0. Cited 3 times in pages 28, 96, and 121.

3GPP. *Service requirements for the 5G system*. 2023. TS22.261,Version 19.2.0. Cited on page 126.

Acatauassu, D. et al. An efficient fronthaul scheme based on coaxial cables for 5G centralized radio access networks. *IEEE Transactions on Communications*, v. 69, n. 2, p. 1343–1357, 2021. Cited on page 124.

Agiwal, M.; Roy, A.; Saxena, N. Next generation 5g wireless networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, v. 18, n. 3, p. 1617–1655, 2016. Cited on page 159.

Anders, G. J.; Silva, A. M. L. da. Cost related reliability measures for power system equipment. *IEEE Trans. Power Sys.*, v. 15, n. 2, p. 654–660, May 2000. Cited 6 times in pages 62, 64, 65, 68, 72, and 83.

Attarifar, M.; Abbasfar, A.; Lozano, A. Modified conjugate beamforming for cell-free massive MIMO. *IEEE Wireless Communications Letters*, v. 8, n. 2, p. 616–619, 2019. Cited on page 22.

Bashar, M.; Cumanan, K.; Burr, A. G.; Ngo, H. Q.; Debbah, M. Cell-free massive MIMO with limited backhaul. In: *2018 IEEE International Conference on Communications (ICC)*. 2018. p. 1–7. Cited 2 times in pages 8 and 111.

Bashar, M. et al. Max–min rate of cell-free massive MIMO uplink with optimal uniform quantization. *IEEE Trans. Wirel. Commun.*, v. 67, n. 10, p. 6796–6815, 2019. Cited 2 times in pages 8 and 38.

Bashar, M. et al. Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization. *IEEE Trans. Wirel. Commun.*, v. 69, n. 1, p. 223–245, 2021. Cited 2 times in pages 38 and 111.

Berghmans, F.; Eve, S.; Held, M. An introduction to reliability of optical components and fiber optic sensors. In: Bock, J. W.; Gannot, I.; Tanev, S. (Ed.). *Optical Waveguide Sensing and Imagingn*. Dordrecht: Springer, 2007, (NAPSB). Chap. 4, p. 73–100. Cited 2 times in pages 88 and 97.

Björnson, E.; Jorswieck, E. *Optimal resource allocation in coordinated multi-cell systems*. : Now Publishers, 2013. (Foundations and Trends ® in Communications and Information). ISBN 9781601986382. Cited 3 times in pages 18, 19, and 29.

Björnson, E.; Hoydis, J.; Sanguinetti, L. Massive MIMO networks: Spectral, energy, and hardware efficiency. *Foundations and Trends ® in Signal Processing*, v. 11, n. 3-4, p. 154–655, 2017. ISSN 1932-8346. Available on: <http://dx.doi.org/10.1561/2000000093>. Cited 15 times in pages 2, 11, 17, 18, 20, 21, 22, 24, 27, 38, 45, 51, 52, 92, and 158.

Björnson, E.; Sanguinetti, L. Making cell-free massive MIMO competitive with MMSE processing and centralized implementation. *IEEE Transactions on Wireless Communications*, v. 19, n. 1, p. 77–90, 2020. Cited 10 times in pages 5, 7, 8, 9, 33, 41, 45, 46, 107, and 128.

Björnson, E.; Sanguinetti, L. Scalable cell-free massive MIMO systems. *IEEE Trans. Commun.*, v. 68, n. 7, p. 4247–4261, Apr. 2020. Cited 18 times in pages 3, 5, 6, 9, 10, 20, 24, 28, 29, 30, 32, 40, 92, 95, 96, 107, 108, and 122.

Bretto, A. *Hypergraph theory: An introduction.* : Springer International Publishing, 2013. (Mathematical Engineering). ISBN 9783319000800. Cited on page 59.

Bussgang, J. J. Crosscorrelation functions of amplitude-distorted gaussian signals. In: . 1952. Available on: <https://api.semanticscholar.org/CorpusID:118091898>. Cited on page 37.

Cable.co.uk. *Worldwide Mobile Data Pricing 2024*. 2024. Accessed: 2024-07-13. Available on: <https://www.cable.co.uk/mobiles/worldwide-data-pricing/>. Cited on page 130.

Checko, A. *Cloud RAN fronthaul: Options, benefits and challenges.* Bremen, Alemanha, 2015. Available on: <https://docplayer.net/41844998-Cloud-ran-fronthaul-options-benefits-and-challenges.html>. Cited on page 159.

Chen, J.; Wosinska, L.; Machuca, C. M.; Jaeger, M. Cost vs. reliability performance study of fiber access network architectures. *IEEE Commun. Mag.*, v. 48, n. 2, p. 56–65, Feb. 2010. ISSN 0163-6804. Cited 2 times in pages 124 and 125.

Cui, Y.; Ingalz, C.; Gao, T.; Heydari, A. Total cost of ownership model for data center technology evaluation. In: *2017 16th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm)*. 2017. p. 936–942. Cited 2 times in pages 116 and 124.

Dahlman, E.; Parkvall, S.; Skold, J. *5G NR: The Next Generation Wireless Access Technology*. Elsevier Science, 2018. ISBN 9780128143230. Available on: <https://books.google.com.br/books?id=lcSLswEACAAJ>. Cited 6 times in pages 161, 162, 163, 164, 165, and 166.

Dahlman, E.; Parkvall, S.; Skold, J. *5G NR: The next generation wireless access technology*. Elsevier Science, 2020. ISBN 9780128223208. Available on: <https://books.google.com.br/books?id=PZH9DwAAQBAJ>. Cited on page 122.

Debaillie, B.; Desset, C.; Louagie, F. A flexible and future-proof power model for cellular base stations. In: *Proc. IEEE VTC Spring*. 2015. p. 1–7. Cited 5 times in pages XX, 42, 43, 113, and 123.

Demir, Ö.; Björnson, E.; Sanguinetti, L. *Foundations of user-centric cell-free massive MIMO*. Hanover, MA, USA: Now Publ., 2021. (Foundations and Trends ® in Signal Processing). ISBN 9781680837902. Cited 28 times in pages 3, 9, 10, 18, 21, 22, 23, 24, 26, 27, 36, 37, 40, 41, 44, 45, 46, 47, 48, 49, 50, 51, 52, 92, 107, 108, 110, and 122.

Demir, O. T.; Masoudi, M.; Björnson, E.; Cavdar, C. Cell-free massive MIMO in O-RAN: Energy-aware joint orchestration of cloud, fronthaul, and radio resources. *IEEE J. Sel. Areas Commun.*, Jan. 2024. Cited 7 times in pages 5, 10, 33, 43, 44, 107, and 113.

Diestel, R. *Graph theory*. 5th. ed. 2017. (Springer Graduate Texts in Mathematics). ISBN 9783961340057. Cited 5 times in pages 54, 55, 56, 57, and 58.

EARTH. *Deliverable D2.3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown*. 2010. Cited on page 113.

El-Saleh, A. A. et al. Measurement analysis and performance evaluation of mobile broadband cellular networks in a populated city. *Alexandria Engineering Journal*, v. 66, p. 927–946, 2023. ISSN 1110-0168. Available on: <https://www.sciencedirect.com/science/article/pii/S1110016822007086>. Cited on page 110.

Ericsson. *Ericsson mobility report june 2021*. 2021. Available on: <https://www.ericsson.com/4a03c2/assets/local/reports-papers/mobility-report/documents/2021/june-2021-ericsson-mobility-report.pdf>. Cited on page 120.

Eriksson, A.; Forsman, M.; Ronkainen, H.; Willars, P.; Ostberg, C. 5g new radio ran and transport choices that minimize tco. *Ericsson Technology Review*, p. 2–11, Nov 2019. ISSN 0014-0171. Cited 2 times in pages 159 and 160.

Europe, F. C. *European FTTH/B market panorama 2023*. 2023. Accessed: Sep 11, 2023. Available on: <https://www.ftthcouncil.eu/knowledge-centre/all-publications-and-assets/1707/european-ftth-b-market-panorama-2023>. Cited on page 118.

Farias, F. et al. Cost and energy-efficient backhaul options for heterogeneous mobile network deployments. *Photon Netw Commun*, v. 32, p. 422–437, 2016. Cited 2 times in pages 117 and 118.

Farias, F. de S. *Designing Cost-Efficient Transport Solutions For Fixed And Mobile Broadband Access Network*. Tese (Doutorado) — UFPA, Belém-PA, Brasil, Mar 2016. Cited 6 times in pages 60, 75, 76, 77, 157, and 159.

Femenias, G.; Riera-Palou, F. Fronthaul-constrained cell-free massive MIMO with low resolution ADCs. *IEEE Access*, v. 8, p. 116195–116215, Jun. 2020. Cited 16 times in pages 5, 9, 33, 36, 37, 38, 39, 40, 41, 45, 46, 47, 48, 101, 107, and 111.

Fernandes, A. L. P. *Technical economic evaluation of alternatives backhaul and fronthaul for centralized radio architectures in 5G indoor settings*. Master thesis in Electrical Engineering — Federal University of Pará, Belém, May 2019. Available at: <http://repositorio.ufpa.br/jspui/handle/2011/11326>. Cited 4 times in pages 11, 60, 65, and 72.

Fernandes, A. L. P.; Farias, F. S.; Ohashi, A. A.; Oliveira, M. V. P.; Costa, J. C. W. A. A techno-economic assessment of 5g networks with passive optical network-based mobile backhaul and hybrid fiber-copper fronthaul. In: *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2019)*. Sociedade Brasileira de Computação, 2019. (SBRC 2019). Available on: <http://dx.doi.org/10.5753/sbrc.2019.7371>. Cited on page 60.

Fernandes, A. L. P.; Souza, D. D.; Costa, D. B. da; Cavalcante, A. M.; Costa, J. C. W. A. Cell-free massive MIMO with segmented fronthaul: Reliability and protection aspects. *IEEE Wireless Communications Letters*, v. 11, n. 8, p. 1580–1584, 2022. Cited on page 124.

Fernandez, A.; Stol, N. CAPEX and OPEX simulation study of cost-efficient protection mechanisms in passive optical networks. *Opt. Switch. Netw.*, Elsevier Science Publishers B. V., Amsterdam, Países Baixos, v. 17, n. C, p. 14–24, Jul 2015. ISSN 1573-4277. Cited 6 times in pages 60, 67, 72, 74, 77, and 124.

Fernandez, A.; Stol, N. Economic, dissatisfaction, and reputation risks of hardware and software failures in PONs. *IEEE/ACM Trans. Netw.*, v. 25, n. 2, p. 1119–1132, Nov. 2016. Cited 7 times in pages 10, 11, 60, 65, 67, 77, and 81.

Fiorani, M. et al. Modeling energy performance of C-RAN with optical transport in 5G network scenarios. *Journal of Optical Communications and Networking*, v. 8, n. 11, p. B21–B34, 2016. Cited on page 118.

Fiorani, M.; Tombaz, S.; Monti, P.; Casoni, M.; Wosinska, L. Green backhauling for rural areas. In: *Proc. ONDM*. 2014. p. 114–119. Cited 2 times in pages 109 and 157.

Freitas, M. M. M. et al. Scalable user-centric distributed massive MIMO systems with limited processing capacity. In: *Proc. IEEE ICC*. 2023. p. 1–7. Cited 6 times in pages 5, 9, 10, 29, 107, and 122.

Frenger, P. et al. *Radio Stripes - Technical Details*. 2019. Ericsson Internal Report. Cited 2 times in pages 88 and 91.

Gesbert, D. et al. Multi-cell MIMO cooperative networks: A new look at interference. *IEEE Journal on Selected Areas in Communications*, v. 28, n. 9, p. 1380–1408, 2010. Cited on page 23.

Goodrich, M.; Tamassia, R. *Algorithm design and applications*. : Wiley, 2014. ISBN 9781119028482. Cited on page 58.

Gray, R.; Davisson, L. *An introduction to statistical signal processing*. : Cambridge University Press, 2004. ISBN 9781139456289. Cited 2 times in pages 38 and 39.

Hardy, D. et al. An analytical framework for estimating tco and exploring data center design space. In: *2013 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*. 2013. p. 54–63. Cited 3 times in pages 116, 117, and 124.

HKT GSA; Huawei. *Indoor 5G scenario oriented white paper*. 2019. Cited 2 times in pages 126 and 130.

Interdonato, G.; Björnson, E.; Ngo, H.; Frenger, P.; Larsson, E. Ubiquitous Cell-Free Massive MIMO Communications. *EURASIP Journal on Wireless Communications and Networking*, p. 1–13, 08 2019. Cited 20 times in pages XVII, 3, 4, 5, 6, 7, 8, 10, 19, 24, 25, 80, 81, 85, 107, 109, 113, 119, 131, and 132.

Interdonato, G.; Frenger, P.; Larsson, E. G. Scalability aspects of cell-free massive MIMO. In: *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*. 2019. p. 1–6. Cited 2 times in pages 7 and 9.

Jahid, A.; Hossain, M. S.; Monju, M. K. H.; Rahman, M. F.; Hossain, M. F. Techno-economic and energy efficiency analysis of optimal power supply solutions for green cellular base stations. *IEEE Access*, v. 8, p. 43776–43795, 2020. Cited 2 times in pages 117 and 124.

Jakopin, N. et al. *The Evolution of Data Growth in Europe*. 2023. Cited on page 130.

Jha, A.; Saha, D. Techno-economic analysis of 5G deployment scenarios involving massive MIMO hetnets over mmwave: A case study on the us state of texas. In: *Proceedings of the 51st Hawaii International Conference on System Sciences*. 2018. Cited on page 11.

Jin, S.-N.; Yue, D.-W.; Nguyen, H. H. Spectral and energy efficiency in cell-free massive MIMO systems over correlated rician fading. *IEEE Systems Journal*, v. 15, n. 2, p. 2822–2833, 2021. Cited on page 28.

Kim, T.; Kim, H.; Choi, S.; Hong, D. How will cell-free systems be deployed? *IEEE Communications Magazine*, v. 60, n. 4, p. 46–51, 2022. Cited 2 times in pages 5 and 10.

Klein, T. *eCPRI Overview*. 2017. Available on: <http://www.ieee802.org/1/files/public/docs2017/cm-mustala-eCPRI-Overview-0917.pdf>. Cited on page 122.

Kletz, T. *Hazop and Hazan: Identifying and Assessing Process Industry Hazards*. : Institution of Chemical Engineers, 2001. ISBN 9780852955062. Cited 2 times in pages 66 and 68.

Larsen, L. M. P.; Checko, A.; Christiansen, H. L. A survey of the functional splits proposed for 5g mobile crosshaul networks. *IEEE Communications Surveys & Tutorials*, v. 21, n. 1, p. 146–172, 2019. Cited 4 times in pages 159, 167, 168, and 169.

Liggett, T. *Continuous Time Markov Processes: An Introduction*. : American Mathematical Society, 2010. (Graduate studies in mathematics). ISBN 9780821849491. Cited 7 times in pages 59, 60, 61, 62, 64, 65, and 68.

Liu, L. et al. Analysis of handover performance improvement in cloud-ran architecture. In: *7th International Conference on Communications and Networking in China*. 2012. p. 850–855. Cited on page 159.

Luxalight. *Datasheet for LS24B120X2835PLX LED Strip*. 2021. <https://www.luxalight.eu/sites/default/files/downloads/2021-10/Datasheet_LS24B120X2835PLX.pdf>. Accessed: 2023-04-23. Cited on page 92.

Mahloo, M.; Monti, P.; Chen, J.; Wosinska, L. Cost modeling of backhaul for mobile networks. In: *2014 IEEE International Conference on Communications Workshops (ICC)*. 2014. p. 397–402. Cited on page 118.

Martino, C. D. et al. Lessons learned from the analysis of system failures at petascale: The case of blue waters. In: *Proc. IEEE/IFIP DSN*. 2014. p. 610–621. Cited on page 124.

Marzetta, T. L. Massive MIMO: An introduction. *Bell Labs Technical Journal*, v. 20, p. 11–22, 2015. Cited on page 2.

Max, J. Quantizing for minimum distortion. *IRE Transactions on Information Theory*, v. 6, n. 1, p. 7–12, 1960. Cited 2 times in pages XX and 40.

Mezghani, A.; Nossek, J. A. Capacity lower bound of MIMO channels with output quantization and correlated noise. In: *Proc. IEEE Int. Symp. Inf. Theory*. Cambridge, MA, USA: , 2012. p. 1–5. Cited on page 39.

Miretti, L.; Björnson, E.; Gesbert, D. Precoding for scalable cell-free massive MIMO with radio stripes. In: *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 2021. p. 411–415. Cited 2 times in pages 8 and 41.

Monti, P.; Tombaz, S.; Wosinska, L.; Zander, J. Mobile backhaul in heterogeneous network deployments: Technology options and power consumption. In: *2012 14th International Conference on Transparent Optical Networks (ICTON)*. 2012. p. 1–7. Cited 2 times in pages 117 and 118.

Nahas, M.; Saadani, A.; Charles, J.-P.; El-Bazzal, Z. Base stations evolution: Toward 4g technology. In: *2012 19th International Conference on Telecommunications (ICT)*. 2012. p. 1–6. Cited on page 159.

NGMN Alliance e.V. *6G Requirements and design considerations*. 2023. Cited on page 126.

Ngo, H. Q.; Ashikhmin, A.; Yang, H.; Larsson, E. G.; Marzetta, T. L. Cell-free massive MIMO: Uniformly great service for everyone. In: *2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. 2015. p. 201–205. Cited on page 7.

Ngo, H. Q.; Ashikhmin, A.; Yang, H.; Larsson, E. G.; Marzetta, T. L. Cell-free massive MIMO versus small cells. *IEEE Transactions on Wireless Communications*, v. 16, n. 3, p. 1834–1850, 2017. Cited 5 times in pages 3, 5, 6, 7, and 107.

Ngo, H. Q.; Tataria, H.; Matthaiou, M.; Jin, S.; Larsson, E. G. On the performance of cell-free massive MIMO in ricean fading. In: *2018 52nd Asilomar Conference on Signals, Systems, and Computers*. 2018. p. 980–984. Cited on page 9.

Ngo, H. Q.; Tran, L.-N.; Duong, T. Q.; Matthaiou, M.; Larsson, E. G. On the total energy efficiency of cell-free massive MIMO. *IEEE Transactions on Green Communications and Networking*, v. 2, n. 1, p. 25–39, 2018. Cited on page 7.

NTT-DOCOMO. *R3-162102: CU-DU split: Refinement for Annex A*. 2016. Cited on page 169.

Oliva, A. D. L. et al. Xhaul: Toward an integrated fronthaul/backhaul architecture in 5g networks. *IEEE Wireless Communications*, v. 22, n. 5, p. 32–40, 2015. Cited on page 159.

Oughton, E. J.; Lehr, W. Surveying 5G techno-economic research to inform the evaluation of 6G wireless technologies. *IEEE Access*, v. 10, p. 25237–25257, 2022. Cited 4 times in pages 11, 12, 75, and 77.

Özdogan, O.; Björnson, E.; Larsson, E. G. Massive MIMO with spatially correlated Rician fading channels. *IEEE Trans. Commun.*, v. 67, n. 5, p. 3234–3250, 2019. Cited 3 times in pages 26, 28, and 121.

Ozdogan, O.; Bjornson, E.; Larsson, E. G. Massive MIMO with spatially correlated rician fading channels. *IEEE Transactions on Communications*, v. 67, n. 5, p. 3234–3250, 2019. Cited on page 28.

Papoulis, A.; Pillai, S. *Probability, random variables, and stochastic processes.* : McGraw-Hill, 2002. (McGraw-Hill series in electrical and computer engineering). ISBN 9780071226615. Cited 3 times in pages 59, 60, and 61.

Patterson, M.; Costello, D.; Grimm, P. *Data center TCO: A comparison of high-density and low-density spaces.* 2007. Available on: <https://citeseerx.ist.psu.edu/documentrepid=rep1&type=pdf&doi=22690d3a390e12a9d38296107b86804fcce7fc64>. Cited on page 116.

Perrin, S. *Evolving to a Open C-RAN architecture for 5G.* Japão, 2017. Available on: <https://www.fujitsu.com/us/Images/FNC-Fujitsu-Evolving-to-an-Open-C-RAN-Architecture-for-5G-White-Paper.pdf>. Cited on page 160.

Philpott, M.; Fellenbaum, A.; Frey, D. *Global fiber development index: 2020.* 2020. Accessed: Sep 11, 2023. Available on: <https://omdia.tech.informa.com/OM014270/Global-Fiber-Development-Index-Analysis-2020>. Cited on page 118.

Pishro-Nik, H. *Introduction to probability, statistics, and random processes.* : Kappa Research, LLC, 2014. ISBN 9780990637202. Cited 2 times in pages 38 and 39.

Polegre, A. A.; Riera-Palou, F.; Femenias, G.; Armada, A. G. Channel hardening in cell-free and user-centric massive MIMO networks with spatially correlated ricean fading. *IEEE Access*, v. 8, p. 139827–139845, 2020. Cited 3 times in pages 5, 9, and 28.

Principled Technologies. *Reduce hands-on deployment times to near zero with iDRAC9 automation.* 2020. Cited on page 124.

Principled Technologies. *The science behind the report: Empower more virtual desktop users with DELL EMC powerEdge R650xs servers with 3rd generation intel xeon scalable processors.* 2021. Cited on page 123.

Pärssinen, M.; Wahlroos, M.; Manner, J.; Syri, S. Waste heat from data centers: An investment analysis. *Sustainable Cities and Society*, v. 44, p. 428–444, 2019. ISSN 2210-6707. Available on: <https://www.sciencedirect.com/science/article/pii/S2210670718314318>. Cited on page 77.

Qureshi, H. N.; Manalastas, M.; Zaidi, S. M. A.; Imran, A.; Kalaa, M. O. A. Service level agreements for 5G and beyond: Overview, challenges and enablers of 5g-healthcare systems. *IEEE Access*, v. 9, p. 1044–1061, 2021. Cited on page 110.

Roy, J. S.; Mishra, S. S. Performance of sc-fdma for lte uplink under different modulation schemes. In: *2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE)*. 2019. p. 202–206. Cited on page 162.

Selim, M.; Kamal, A. E. Post-disaster 4G/5G network rehabilitation using drones: Solving battery and backhaul issues. In: *2018 IEEE Globecom Workshops (GC Wkshps)*. 2018. p. 1–6. Cited 3 times in pages 4, 10, and 81.

Selim, M.; Kamal, A. E.; Elsayed, K.; Abdel-Atty, H. M.; Alnuem, M. Fronthaul cell outage compensation for 5G networks. *IEEE Commun. Mag.*, v. 54, n. 8, p. 169–175, Aug. 2016. Cited 3 times in pages 4, 10, and 81.

Shaik, Z. H.; Björnson, E.; Larsson, E. MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes. 12 2020. Cited 5 times in pages 8, 41, 80, 81, and 85.

Sigwele, T.; Alam, A. S.; Pillai, P.; Hu, Y. F. Energy-efficient cloud radio access networks by cloud based workload consolidation for 5G. *Journal of Network and Computer Applications*, v. 78, p. 1–8, 2017. ISSN 1084-8045. Available on: <https://www.sciencedirect.com/science/article/pii/S1084804516302740>. Cited 2 times in pages XVI and 115.

Souza, D. D. et al. Effective channel blind estimation in cell-free massive MIMO networks. *IEEE Wireless Communications Letters*, v. 11, n. 3, p. 468–472, 2022. Cited 2 times in pages 22 and 46.

Souza, M. A. D. et al. A techno-economic framework for installing broadband networks in rural and remote areas. *IEEE Access*, v. 9, p. 58421–58447, 2021. Cited on page 76.

Tombaz, S. et al. Impact of backhauling power consumption on the deployment of heterogeneous mobile networks. In: *2011 IEEE Global Telecommunications Conference - GLOBECOM 2011*. 2011. p. 1–5. Cited 2 times in pages 157 and 158.

Tools, G. *5G NR Throughput Calculator*. 2024. <https://5g-tools.com/5g-nr-throughput-calculator/>. Accessed: February 9, 2024. Cited on page 169.

Trinh, H. D.; Bui, N.; Widmer, J.; Giupponi, L.; Dini, P. Analysis and modeling of mobile traffic using real traces. In: *Proc. IEEE PIMRC*. 2017. p. 1–6. Cited on page 120.

Tukmanov, A. et al. Fronthauling for 5g and beyond. In: . 2017. Cited on page 159.

Udalcovs, A. et al. Total cost of ownership of digital vs. analog radio-over-fiber architectures for 5G fronthauling. *IEEE Access*, v. 8, p. 223562–223573, 2020. Cited on page 120.

Xiao, Y.; Mähönen, P.; Simić, L. System cost analysis of scalable cell-free massive MIMO architectures for 6G networks. In: *2022 IEEE Globecom Workshops (GC Wkshps)*. 2022. p. 310–316. Cited 2 times in pages 5 and 12.

Xiao, Y.; Mähönen, P.; Simić, L. Energy and economic efficiency of scalable cell-free massive MIMO networks. In: *2023 IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. 2023. p. 1–6. Cited 2 times in pages 5 and 12.

Yaghoubi, F. et al. A techno-economic framework for 5G transport networks. *IEEE Wireless Communications*, v. 25, n. 5, p. 56–63, 2018. Cited 6 times in pages 12, 75, 76, 77, 117, and 118.

Zhang, J. et al. Prospective multiple antenna technologies for beyond 5G. *IEEE J. on Sel. Areas Commun.*, v. 38, n. 8, p. 1637–1660, Jun. 2020. Cited 3 times in pages 2, 3, and 11.

Zhang, J.; Zhang, J.; Ai, B. Cell-free massive MIMO with low-resolution ADCs over spatially correlated channels. In: *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*. 2020. p. 1–7. Cited on page 45.

Zhang, Y.; Liu, J. Prediction of overall energy consumption of data centers in different locations. *Sensors*, v. 22, n. 10, 2022. ISSN 1424-8220. Available on: <https://www.mdpi.com/1424-8220/22/10/3704>. Cited on page 116.

Zhang, Y.; Zhou, M.; Qiao, X.; Cao, H.; Yang, L. On the performance of cell-free massive MIMO with low-resolution ADCs. *IEEE Access*, v. 7, p. 117968–117977, 2019. Cited on page 38.

Appendix A – Current technologies for cellular networks

This section expands the discussion in Section 2.1 by providing a comprehensive overview of the current technologies used in cellular mobile networks. While these concepts may not be central to the thesis, it is clear that the transition toward cell-free Massive Multiple-Input Multiple-Output (mMIMO) networks will occur through existing cellular network infrastructures. Therefore, understanding the technologies utilized in present-day, 4G and 5G cellular networks, is essential to quickly develop feasible cell-free solutions. Additionally, it's important to note that even in the future where cell-free networks are widely used to provide uniform coverage, cellular hotspots will still be needed to offload localized traffic. This indicates that even as cell-free networks become more popular, coexistence with cellular networks will continue. Lastly, certain concepts from current networks, such as the Orthogonal Frequency-Division Multiplexing (OFDM) modulation used in Long Term Evolution (LTE) and New Radio (NR) standards, will persist in cell-free networks.

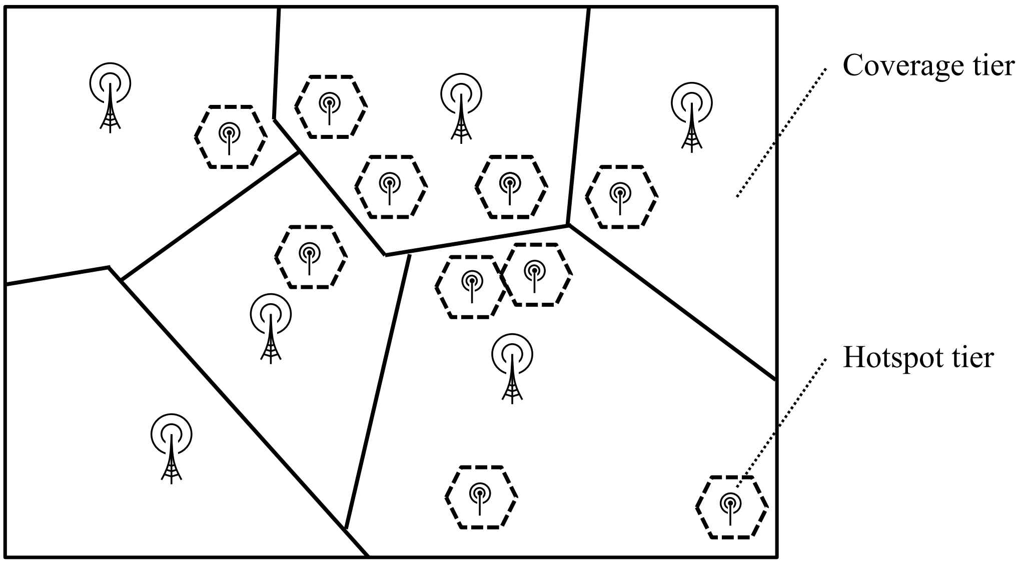
In this context, the study of current technologies for cellular networks is of great value for future developments in cell-free mMIMO.

A.1 HetNets: Coverage and hotspot tiers

The number of cells in a cellular network is constrained by either coverage or capacity considerations. In the first case, the number of cells is determined to ensure an acceptable Signal-to-Noise Ratio (SNR) level across the network. In simple terms, each cell uses the maximum coverage radius that its equipment allows, and the network area is divided into cells based on its individual cell coverage area. On the other hand, in scenarios where capacity is the limiting factor, each cell's coverage radius is extended only to support the maximum demand, constrained by channel or network resources available at the Base Station (BS). Cell densification may be necessary in such cases to accommodate higher traffic demands. This densification involves increasing the number of cells in the network to effectively manage and distribute the traffic load, ensuring that users receive satisfactory service quality even during peak usage periods (Farias, 2016) (Fiorani et al., 2014)(Tombaz et al., 2011).

Cellular networks may be homogeneous, distributing BSs under just a coverage tier, or heterogeneous, distributing BSs under coverage and hotspot tiers. There may be a coverage overlap of cells in different tiers, but not in cells on the same tier. In this way, each User Equipment (UE) is connected to only one of the BSs, which provides its downlink and uplink services. The downlink refers to the signals sent from the BSs to their respective UEs, while the uplink refers to transmissions from the UEs to their respective BSs (Björnson; Hoydis; Sanguinetti, 2017)(Tombaz et al., 2011). Figure 62 exemplifies a mobile cellular network.

Figure 62 – Example of a heterogeneous cellular network. The network coverage area is subdivided into cells that operate individually. The coverage tier seeks to serve all users at any place and anytime. While, the hotspot tier offers high throughput in small local areas to a few UEs.



Source: elaborated by the author.

The coverage tier ideally seeks to serve all users over the large area served by the cellular network at any place and anytime. It generally consists of large outdoor BSs that provide wide-area coverage mobility support and are shared between many UEs. The hotspot tier acts as a traffic offload to specific areas with high data traffic demands, avoiding an overload of the coverage tier. It generally consists of small indoor/outdoor BSs that offer increased throughput in small local areas to a few UEs. If two tiers share the same frequency spectrum, inter-tier coordination is necessary to avoid interference. However, to avoid this extra processing and overhead, it is common to use different spectrums for coverage and hotspots tiers. For example, the coverage tier might use LTE and operate in the 2.1 GHz band, while the hotspot tier might use Wireless Fidelity (Wi-Fi) in the 5 GHz band (Björnson; Hoydis; Sanguinetti, 2017).

The primary advantage of employing two tiers is the ability to offload localized traffic from the hotspot tier, thereby circumventing the need for network densification. This strategy mitigates the requirement for installing additional, potentially unnecessary large base stations, which are associated with high energy consumption and installation costs.

In contrast, the small cells in the hotspot tier are much cheaper and consume way less energy, contributing to overall network efficiency and sustainability (Farias, 2016)(Agiwal; Roy; Saxena, 2016).

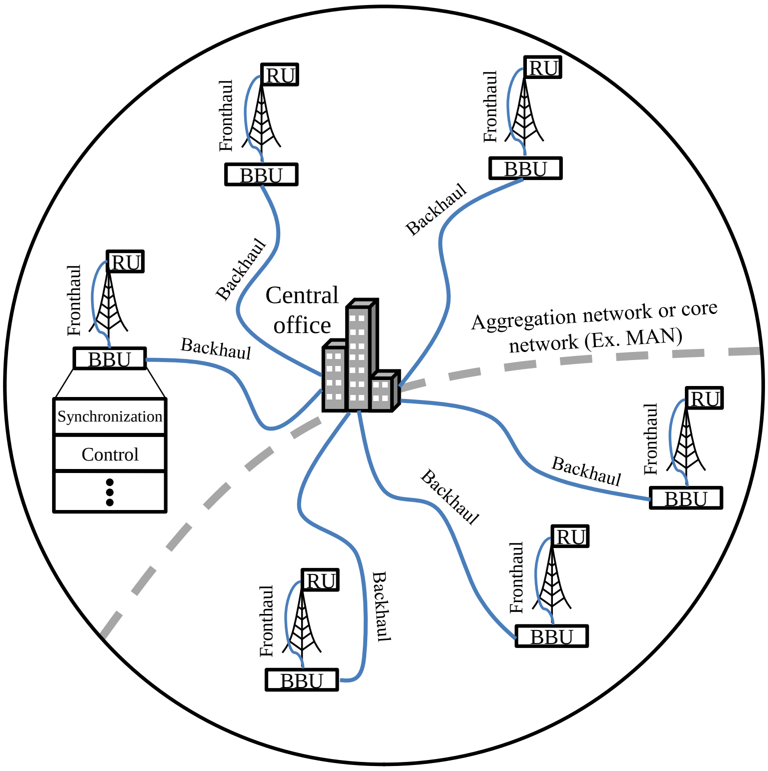
A.2 Cellular RAN architectures

Radio Access Networks (RANs) are composed of passive antennas, Radio Frequency (RF) equipment, digital processors, communication links, and other devices in the BS or connected to it. The set of communication links inside a RAN is called fronthaul and the connection between the RANs and the operator's core network, such as a Metropolitan Area Network (MAN), is called backhaul and traditionally is part of the transport network (Nahas et al., 2012). However, the boundaries between traditional backhaul and fronthaul have become somewhat ambiguous, and the joining of these two concepts is sometimes called x-haul or crosshaul (Oliva et al., 2015)(Larsen; Checko; Christiansen, 2019)(Checko, 2015).

In Fourth Generation of Mobile Systems (4G) cellular mobile systems, a Distributed Radio Access Network (D-RAN) architecture was favored, exemplified by Figure 63. The composition of this architecture involves three main components in each cell site: Baseband Unit (BBU), Radio Unit (RU), and fronthaul links. The first performs the heavy digital signal processing part, including synchronization, control, transport, and baseband functions. The second performs all RF processing functions. Finally, the fronthaul links connect the BBUs to Radio Remote Heads (RRHs) using technologies such as digital Radio-Over-Fiber (RoF) or Radio-Over-Copper (RoC). This architectural design has significantly enhanced BS energy efficiency and streamlined implementation compared to earlier cellular systems, such as Second Generation of Mobile Systems (2G), which relied on more equipment and lengthy copper cables to deliver analog signals to passive antennas, requiring substantial amplification. These improvements have solidified D-RAN as the prevailing RAN architecture in modern mobile communication systems (Checko, 2015)(Eriksson et al., 2019).

Despite being widely implemented, the D-RAN architecture has some disadvantages: (1) the processing infrastructure is rigid, not being able to adapt to the daily movement of users, i.e., in some BSs, there will be underutilization of resources, and in others, there will be a shortage of resources; (2) as there is no joint processing between BSs it becomes more complicated to apply interference cancellation techniques between BSs; and (3) the transition of a user from the coverage area of one BS to another, a procedure known as handover, can generate large quantities of overhead and be a bottleneck for the network, especially in scenarios with a dense deployment of BSs (Tukmanov et al., 2017)(Liu et al., 2012).

Figure 63 – Example of the D-RAN architecture. BBUs are located at cell-site, and backhaul links connect the cells to the operator's Central Office (CO) and core network. All digital processing is done at the cell location.



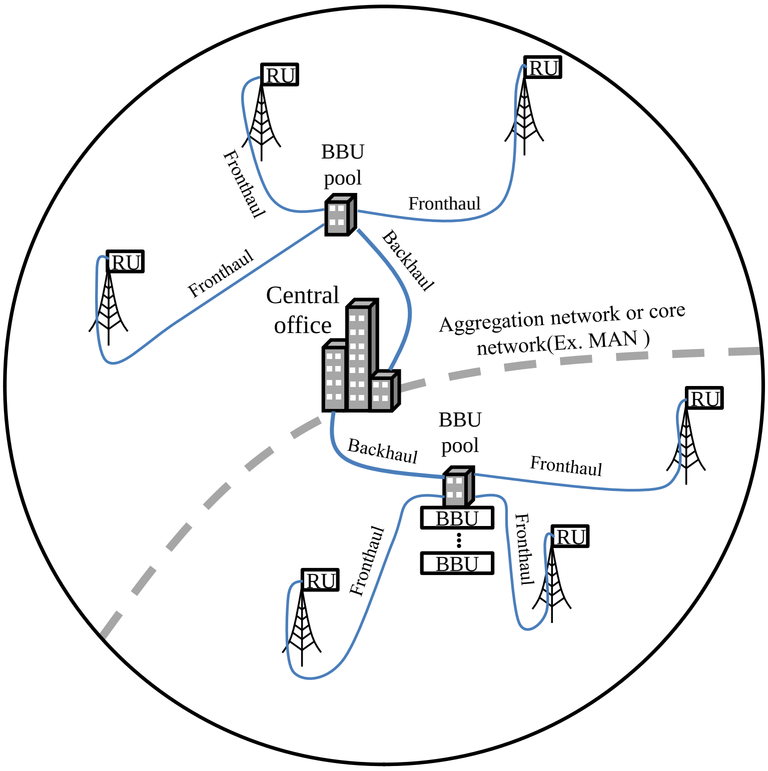
Source: elaborated by the author.

Most disadvantages of D-RAN are solved with the Centralized Radio Access Network (C-RAN) architecture, exemplified in Figure 64, where it is noticed that only the RU are in the BS location. The BBUs of multiple BSs are co-located and work together in a BBU pool, which connects to BS/Transmission-Reception Point (TRP) by fronthaul links and to the operator's CO by backhaul links. Simply put, the great advantage of the C-RAN is that the centralization of digital signal processing facilitates the coordination of radio resources between the BSs/TRPs. Another advantage of the C-RAN architecture is its easier maintenance. The cell site has less equipment than D-RAN, and the centralization of equipment and resources reduces the need for travel to make repairs in the cell site. The concept of network centralization can be raised even further with the virtualization of BBU functions in a set of servers serving a large number of cells, such technology being known as Cloud RAN(Eriksson et al., 2019)(Perrin, 2017).

A.3 4G and 5G wireless access technologies: LTE and NR

The global success of mobile communication was only possible due to multinational agreements on the adopted specifications and standards. This fact allowed interoperability between the products of different vendors or manufacturers and enabled devices and

Figure 64 – Example of the C-RAN architecture. Multiple BBUs are located at a BBU pool. Digital processing resources are shared between cells. Fronthaul links connect the cells to the BBU pools. Backhaul links connect BBU pools to the operator's CO and core network.



Source: elaborated by the author.

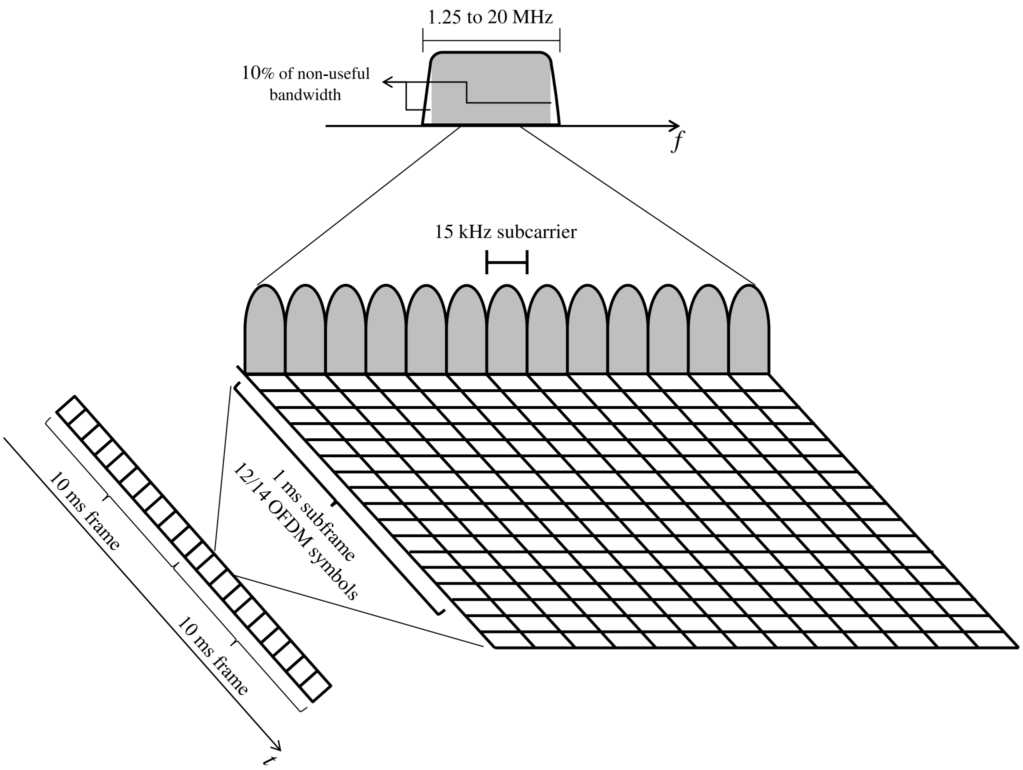
subscriptions to operate on a global basis. The proper global standardization of cellular mobile communication came with the specification of third-generation systems when different regional standardization organizations got together and jointly created the Third Generation Partnership Project (3GPP) to finalize the development of Third Generation of Mobile Systems (3G) technology. Currently, 3GPP is the only significant organization developing technical specifications for cellular mobile communication. Including the LTE for 4G and NR for Fifth Generation of Mobile Systems (5G) (Dahlman; Parkvall; Skold, 2018).

A.3.1 LTE

The LTE standard served as the foundation for 4G radio technology and was established in 3GPP release 8, supporting bandwidths ranging from 1.25 MHz to 20 MHz. The downlink physical communication structure, depicted in Figure 65, is based on Orthogonal Frequency-Division Multiple Access (OFDMA) with a subcarrier spacing of 15 kHz. Despite this, only 90% of the available bandwidth is utilized, resulting in a maximum of 1200 subcarriers. The standard organizes transmissions into 10 ms frames in the time domain, each comprising ten 1 ms subframes. These subframes are further divided into two 0.5 ms slots, each containing 7 or 6 OFDM symbols depending on two

Cyclic Prefixs (CPs) duration options. The first is the normal CP lasting 4.7 μs, and the second is the extended CP lasting 16.67 μs. These two CP options allow for Inter-Symbol Interference (ISI) mitigation, albeit at the expense of transmitting less desired information per time-domain frame. This trade-off is particularly beneficial in scenarios characterized by higher delay spread. Finally, the smallest unit of resources that can be allocated to a user is the Resource Block (RB), which is defined by the band of twelve subcarriers and the duration of one slot (Dahlman; Parkvall; Skold, 2018).

Figure 65 – LTE downlink physical resource grid. Each subframe contains two 0.5 ms slots and consequently can contain 12 or 14 OFDM symbols for CPs with duration of 4.7 μs and 16.67 μs, respectively.



Source: elaborated by the author.

For uplink operation, Single-Carrier Frequency-Division Multiple Access (SC-FDMA) is employed instead of OFDMA, primarily due to the latter's high peak-to-average power ratio, which results in excessive power consumption for signal generation, undesirable for devices with limited battery power like mobile phones. The transmission structure for the SC-FDMA uplink remains mostly similar to the OFDMA downlink one, with the main difference being that all subcarriers in an RB transmit the same symbols, emulating a larger bandwidth single carrier transmission, thereby reducing the symbol time (Roy; Mishra, 2019)(Dahlman; Parkvall; Skold, 2018).

In 3GPP release 8, the standard could operate with paired spectrum using Frequency Division Duplex (FDD) from frequencies below 1 GHz up to about 3 GHz. Additionally, spatial multiplexing support was done only on the downlink using Multiple-Input Multiple-Output (MIMO), and the number of transmission layers was mapped to up to four antennas. Thus, in 3GPP release 8 for two transmission layers LTE, peak data rates of up to 150 and 75 Mbps could be achieved in the downlink and uplink, respectively. Subsequent to Release 8, 3GPP Release 9 introduced unpaired Time Division Duplex (TDD) operation (Dahlman; Parkvall; Skold, 2018).

A.3.2 LTE Evolutions

3GPP releases 8 and 9 form the foundation of LTE. Despite this, the standard was always intended to be upgraded by introducing new technologies and signal processing techniques while maintaining retro-compatibility with the initial release. In this context, 3GPP releases 10 to 15 continuously evolved the LTE to a more capable standard. The initial LTE release was not compliant with International Mobile Telecommunications (IMT)-Advanced requirements defined by International Telecommunication Union - Radiocommunication Sector (ITU-R) for 4G networks, sometimes being even considered a 3.9G system. Only after release 10, the LTE become a 4G system compliant with IMT-Advanced requirements receiving the denomination LTE-advanced (Dahlman; Parkvall; Skold, 2018).

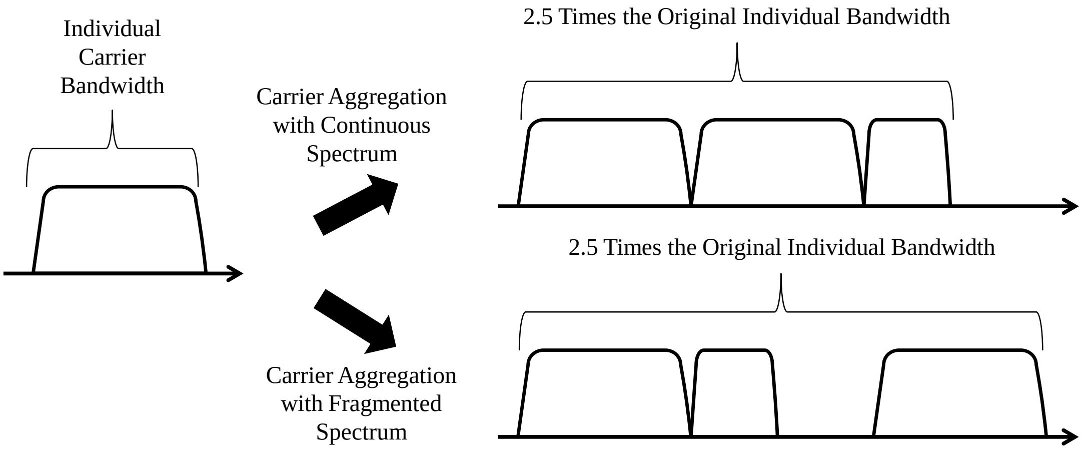
These later evolutions of LTE enabled operation in frequency bands up to 6 GHz, enabling greater spectrum flexibility and peak data rate improvements. Several factors influenced the data rate increase, including increasing the maximum number of transmission layers to 8 and using spatial multiplexing in the uplink. Despite this, two new technologies brought with them the possibility of much higher transmission rates and Quality-of-Service (QoS), Carrier Aggregation (CA) and Coordinated MultiPoint (CoMP) (Dahlman; Parkvall; Skold, 2018).

CA technology allows bands in different carrier frequencies, called carrier components, to be aggregated and utilized together for the transmission and reception of UE signals. Figure 66 exemplifies the concept, demonstrating the bandwidth available for a group of users in one carrier increasing by 2.5 times. This is achieved by aggregating two additional carriers, one with the same amount of bandwidth as the original carrier and another with half. The illustration depicts two aggregation options: continuous or fragmented across the spectrum. The latter approach is particularly beneficial for providing high bandwidth to users in sub-6 GHz frequencies, as this region supports many services, resulting in a highly fragmented spectrum and limited continuous bandwidth resources (Dahlman; Parkvall; Skold, 2018).

The LTE is limited by its design to a maximum band of 20 MHz, and all its processing is projected for this maximum band. CA is an intelligent way to increase bandwidth without breaking the maximum band of the LTE. This also means that compatibility with earlier versions of LTE is guaranteed because each carrier component uses the bandwidth of 3GPP Release 8 framework. Thus, a user with older LTE hardware can connect to later releases, the disadvantage in this case is that the user with older hardware can not take advantage of the CA (Dahlman; Parkvall; Skold, 2018).

In simpler terms, it is like allowing the user to use multiple bands to attain a higher bandwidth than 20 MHz. Initially, up to five carrier components with equal or different bandwidths could be aggregated. Later, this value was increased to 32, resulting in a total bandwidth of 640 MHz (Dahlman; Parkvall; Skold, 2018).

Figure 66 – Example of CA use cases for continuous and fragmented spectrum.



Source: elaborated by the author.

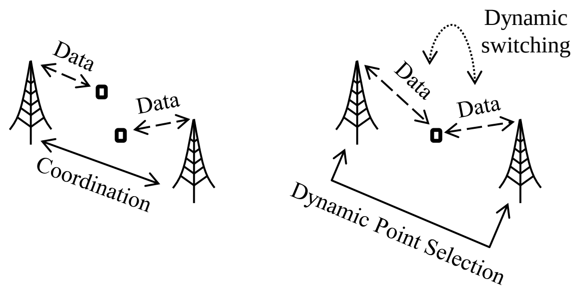
CoMP allows multipoint coordination and transmission. It can be implemented in three ways on LTE, exemplified in Figure 67. In the first, data transmission to the UE is done only from a specific TRP, but the planning and adaptation of the radio link are coordinated between several TRPs. In the second case, data transmission to a device can be performed from multiple TRPs, in such a way that the user can dynamically select a TRP. In the last case, a user can be jointly served by multiple TRPs. One important caveat about CoMP in current 3GPP standards is that the joint transmission is implemented by considering different data streams from each TRP. This is an important difference from cell-free systems, which transmit the same data in a coherent way from all coordinating TRPs (Dahlman; Parkvall; Skold, 2018).

Nowadays LTE can support data rates of more than 3 Gbps, mainly due the CA and the utilization of higher level modulations, such as 256-Quadrature Amplitude Modulation (QAM) (Dahlman; Parkvall; Skold, 2018).

A.3.3 5G NR

Despite the LTE being a very capable standard, there are requirements for future networks that are not possible for LTE to support even with new evolutions. Some new efficient technical telecommunication solutions are not compatible with the basic

Figure 67 – Example of the three ways of CoMP implementation considered on LTE.



Source: elaborated by the author.

inter-workings of the initial release of LTE. In this context, for 5G 3GPP started the development of a new radio access technology known as NR (Dahlman; Parkvall; Skold, 2018).

The NR has a downlink transmission scheme based on OFDMA and an uplink transmission scheme based on OFDMA or SC-FDMA. Moreover, the standard supports a flexible OFDM numerology with subcarrier spacings of 15 kHz to 240 kHz in two operational frequency ranges, the first in the same sub-6GHz region where LTE operated (0.45 - 6 GHz) and the second at Millimeter-Waves (mmWaves) (24.25 - 52.6 GHz). The possible increased subcarrier spacing concerning LTE is necessary for the system to handle the increased phase noise at the second frequency range. Finally, NR supports up to 100 MHz and 400 MHz bandwidths for the frequency ranges 1 and 2, respectively, supporting up to 3300 sub-carriers. In this way, NR supports a wide range of scenarios, from sub-1 GHz implementations to millimeter waves with wide spectrum allocation operations (Dahlman; Parkvall; Skold, 2018).

Furthermore, CA enables the standard to support higher bandwidths. In terms of multiple antenna support, spatial multiplexing is natively supported on both uplink and downlink, and up to 8 transmission layers can be adopted, and analog beamforming is supported. For duplex, the frequency range 1 operates under paired and unpaired spectrum using FDD and TDD. Despite this, the frequency range 2 only operates with unpaired spectrum using TDD. Table 14 summarizes various NR parameters for the two frequency ranges. It is possible to note that not all subcarriers spacing are available for any carrier frequency. The subcarrier spacing of 60 kHz is possible for both frequency ranges. However, smaller subcarrier spacings are used only on frequency range 1, and bigger subcarrier spacings are used only on frequency range 2 (Dahlman; Parkvall; Skold, 2018).

The NR time-domain structure is similar to that of LTE, with 10 ms frames divided into ten subframes of 1 ms duration, each divided into slots consisting of 14 OFDM symbols each. However, unlike LTE, NR does not necessarily use two slots of 0.5 ms per subframe. The duration of a slot depends on the spacing between the subcarriers. For 15

Table 14 – NR possible transmission parameters overview.

5G NR parameter	**Frequency range 1 (0.45 - 6 GHz)**	**Frequency range 2 (24.25 - 52.6 GHz)**
Bandwidth options (MHz)	5, 10 , 15, 20, 25, 30, 40, 50, 60, 70, 80, 90 or 100	50, 100, 200, 300, or 400
Subcarrier spacing options (kHz)	15, 30 or 60	60, 120 or 240
Maximum number of subcarriers	3300	
Maximum number of CA	16	
Duplex	TDD or FDD	TDD

kHz, the duration is 1 ms, in such a way that the number of transmitted OFDM symbols per frame is equal in NR and LTE when the latter uses the normal CP. For the other subcarrier spacings, each doubling concerning 15 kHz halves the slot time, so that for 240 kHz, the slot duration is only 0.0625 ms. This reduction of slot duration can be advantageous for reducing latency since scheduling happens on a slot base. Finally, to provide an even more efficient approach to low latency transmissions, NR allows transmissions in a fraction of a slot, sometimes referred to as "mini-slot" transmission. These transmissions can also pre-empty an already-in-progress slot-based transmission to another device, allowing for immediate data transmission that requires very low latency. Table 15 presents the duration in time of different parameters depending on the subcarrier spacing (Dahlman; Parkvall; Skold, 2018).

Table 15 – Subcarrier spacings supported by NR.

Subcarrier spacing (kHz)	**Slot time (ms)**	**Useful symbol time (μs)**	**Cyclic prefix time (μs)**
15	1	66.7	4.7
30	0.5	33.3	2.3
60	0.25	16.7	1.2
120	0.125	8.33	0.59
240	0.0625	4.17	0.29

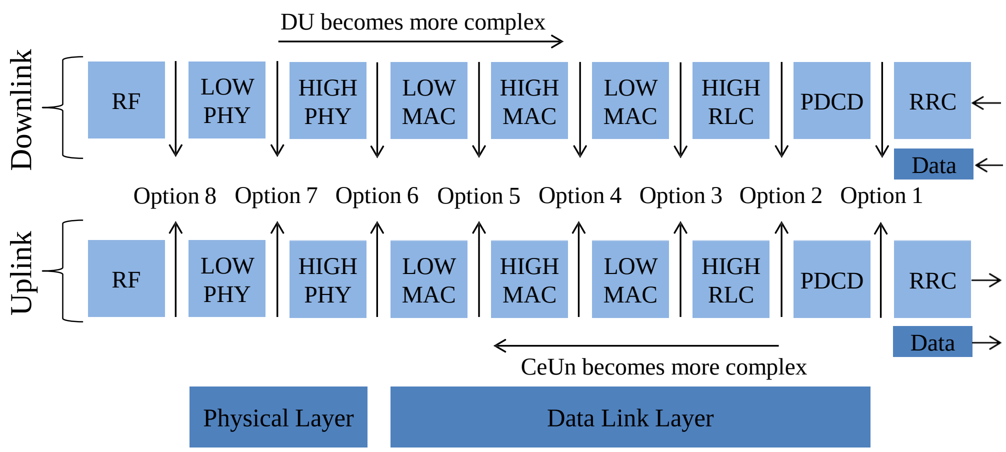
Based on the CP duration presented in Table 15, it is possible to note that NR operates with 14 OFDM symbols per slot. However, there may be delay spread sensitive scenarios for both frequency ranges. For this reason, the subcarrier spacing of 60 kHz can operate with an extended CP in the same way as LTE, allowing only 12 OFDM symbols per slot (Dahlman; Parkvall; Skold, 2018).

A.4 3GPP functional splitting options for cellular networks

Moving digital processing tasks from the cells to a centralized location, as discussed with the C-RAN architecture for cellular networks, offers improved resource management. However, limiting the functions at the BS to analog radio processing and relying on digital communication between the BBU and the RU significantly increases capacity demand on the fronthaul network. A more effective approach involves incorporating additional digital processing functions locally at the cell site to alleviate the overhead of processed data. This strategy can achieve a better balance between resource management and fronthaul requirements, enhancing the overall efficiency of the cellular network architecture (Larsen; Checko; Christiansen, 2019).

Several different functional splits are currently under investigation by 3GPP for use in NR, presented in Figure 68. These split options involve segmenting the user plane protocol sublayers and sometimes their associated functions between the Central Unit (CeUn) and the Distributed Unit (DU). In some scenarios, two splits can be employed simultaneously, with one option implemented between the CeUn and DU, and a higher-level option applied in the link between the DU and RU. In this latter case, RF functions are always located at the RU. Otherwise, they reside on the DU. Finally, functions related to Radio Remote Control (RRC) are always performed by the CeUn. From now on, only the division between CeUn and DU is considered, but it is important to remember that a second division is possible (Larsen; Checko; Christiansen, 2019).

Figure 68 – Schematic of the functional splitting options defined by 3GPP for NR.



Source: elaborated by the author.

Option 8 is the traditional concept of C-RAN, with all digital baseband processing on the CeUn and just the analog RF processing on the DU, and Option 1 is equivalent to the D-RAN architecture, having all processing in cell site. Option 2 lets the physical layer and most of the data link layer in the DU, only Packet Data Convergence Protocol (PDCP) functions are performed centralized. These include Internet Protocol (IP) header compression to reduce the number of bits transmitted over the radio interface, ciphering to protect against eavesdropping, and integrity protection to ensure that control messages originate from the correct source (Larsen; Checko; Christiansen, 2019).

Options 3 and 4 partially and fully centralize Radio Link Control (RLC) sublayer, respectively, which is responsible for the segmentation, i.e., fitting the received data in the adequate transport block size, and retransmission of erroneous packets coming from lower layers. Options 1 to 4 keep the Medium Access Control (MAC) sublayer entirely on the DU, which is responsible for multiplexing of logical channels, Hybrid Automatic Repeat Request (HARQ) retransmissions, and scheduling and scheduling-related functions. In this way, these options may not be desirable in a coordinated system with shared resources between BS. Despite this, they may be applicable in scenarios with limited fronthaul bandwidth (Larsen; Checko; Christiansen, 2019).

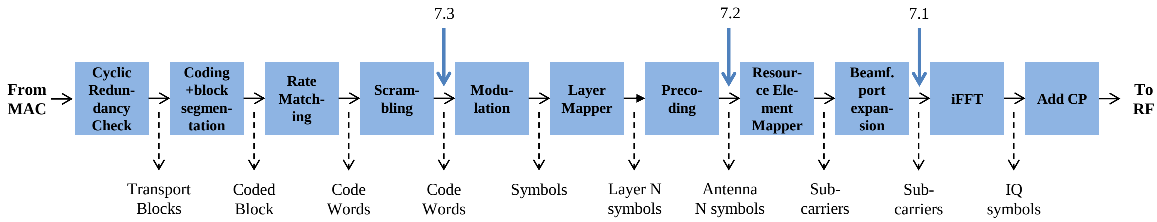
Option 5 centralizes the overall scheduler of the MAC sublayer in the CeUn but places the remaining low-MAC layer functions in each DU, which will handle time-critical procedures in the HARQ locally. This fact reduces delay requirements on the fronthaul and enables large length links from DUs to CeUn-pool. The CeUn communication with the DUs happens through scheduling commands, and HARQ reports. A disadvantage of this functional split is that, although the scheduling is centralized, much of the processing has to be performed locally, limiting the benefits of shared processing. Another disadvantage is that there may be limitations in intercell interference mitigation, which will harm CoMP performance (Larsen; Checko; Christiansen, 2019).

Option 6 centralizes all MAC functions and lets all physical processing be handled locally. In this way, neither MAC nor Physical layer (PHY) sublayers have their processing functions separated between CeUn and DU. This eliminates the need for tight interworking between CeUn and DU. Besides that, intercell interference mitigation and CoMP is facilitated while cell load-dependent fronthaul bitrate is provided. Despite this, only 20% of the overall baseband processing occurs outside the PHY sublayer, i.e., most of the processing is being done locally at the cell site. Moreover, this split has very strict delay requirements as the HARQ, and other time-critical procedures are centralized (Larsen; Checko; Christiansen, 2019).

Option 7 divides the PHY functions between CeUn and DU. Since 80% of the baseband processing happens in PHY, there are multiple ways to divide PHY functions between CeUn and DU. 3GPP defines three points of division, illustrated in Figure 69 called options 7.1, 7.2, and 7.3, where the last is only considered by 3GPP in downlink (Larsen; Checko; Christiansen, 2019).

Option 7.3 lets most PHY functions be on the DU, including modulation, MIMO layer, and resource element mapping. After modulation, several bits, depending on the modulation order, are assigned to each symbol. In this way, the fronthaul transports codewords and will have a reduced bitrate compared to the other alternatives of option 7. Besides that, the load on the fronthaul link is cell load-dependent, and Forward Error Correction (FEC) is inside the CeUn, which benefits the close cooperation between the FEC and the MAC layer. Centralized scheduling is possible, but CoMP may be limited due to the potential latencies over the fronthaul network. Despite this, a problem of this

Figure 69 – Option 7 Functional splits in the PHY layer illustrating the exact location of the functional splits proposed by 3GPP.



Source: elaborated by the author.

option is that it can be only used in downlink (Larsen; Checko; Christiansen, 2019).

Option 7.2 adds the modulation, layer mapping and precoding to the CeUn in relation to 7.3. This allows both downlink and uplink to be supported. The fronthaul link transports subframe symbols, as the Fast Fourier Transform (FFT) and resource element mapper are included in the DU, the load on the fronthaul link is cell load-dependent (NTT-DOCOMO, 2016)(Larsen; Checko; Christiansen, 2019).

Option 7.1 lets only FFT and CP add functions of PHY in the DU. The data to be transmitted over the fronthaul interface is represented by subcarriers. By removing the CP and transforming the received signal to frequency-domain using the FFT, guard subcarriers can be removed in the transmission for the DU, significantly reducing fronthaul bitrate compared to split option 8. However, the fronthaul bitrate is constant and not cell-load dependent as the resource element mapping is executed in the CeUn. Centralized scheduling and CoMP can be supported without performance degradation (NTT-DOCOMO, 2016)(Larsen; Checko; Christiansen, 2019).

Table 16 provides an example of the allowable one-way latency and fronthaul bitrate requisites across various 3GPP functional splitting options. The presented values are compatible with NR communication system employing 32 antennas, FDD, numerology of 30 kHz, a bandwidth of 100 MHz, 8 MIMO layers, and modulation of 256-QAM and 64-QAM for downlink and uplink, respectively.

The configuration used in Table 16 yield peak user data rates of 4674 Mbps for the downlink and 3750 Mbps for the uplink transmissions (Tools, 2024). The calculation of the required bitrate was based on equations from NTT-DOCOMO (2016), Larsen, Checko and Christiansen (2019), and 3GPP (2012), entailing the scaling of peak rates and bit width manipulation, and a sampling rate of 122.88 MS/s, compatible with the considered bandwidth and the Enhanced Common Public Radio Interface (eCPRI) protocol.

Table 16 – Example of requisites across various 3GPP functional splitting options. Values are compatible with a NR communication system employing 32 antennas, FDD, numerology of 30 kHz, a bandwidth of 100 MHz, 8 MIMO layers, and modulation of 256-QAM and 64-QAM for downlink and uplink.

Functional split option	**Downlink fronthaul bitrate (Gbps)**	**Uplink fronthaul bitrate (Gbps)**	**Allowed one-way latency (ms)**
1	4.67	3.75	10
2	4.69	3.77	1.5-10
3	4.67	3.75	1.5-10
4	4.67	3.75	0.1
5	4.81	6.21	0.1
6	4.81	6.21	0.25
7.3	4.81	-	0.25
7.2	11.06-24.37	14.90-23.77	0.25
7.1	41.52-94.74	59.22-94.70	0.25
8	55.05-125.83	78.64-125.83	0.25