CHALMERS
UNIVERSITY OF TECHNOLOGY

A Cost Assessment Methodology for User-Centric Distributed Massive MIMO Architectures

Citation for the original published paper (version of record):

Fernandes, A., Souza, D., Natalino Da Silva, C. et al (2024). A Cost Assessment Methodology for User-Centric Distributed Massive MIMO Architectures. IEEE Open Journal of the Communications Society, 5: 3517-3543. http://dx.doi.org/10.1109/OJCOMS.2024.3406374

IEEE Open Journal of the Communications Society

Received 12 February 2024; revised 15 April 2024; accepted 19 May 2024; Date of publication 29 May 2024; date of current version August 5, 2024.

Digital Object Identifier 10.1109/10.1109/OJCOMS.2024.3406374

A Cost Assessment Methodology for User-Centric Distributed Massive MIMO Architectures

ANDRÉ L. P. FERNANDES[1], DAYNARA D. SOUZA[12], CARLOS NATALINO[3] (Member, IEEE), FEDERICO TONINI[4] (Member, IEEE), ANDRÉ M. CAVALCANTE[5] (Member, IEEE), PAOLO MONTI[3] (Senior Member, IEEE), AND JOÃO C. W. A. COSTA[1] (Senior Member, IEEE)

[1]Applied Electromagnetism Laboratory (LEA), Federal University of Pará (UFPA), Belém, 66075-110 Brazil
[2]Lappeenranta-Lahti University of Technology, Lappeenranta, 53850 Finland
[3]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, 412 96 Sweden
[4]Wireless Networking Laboratory (WiLab), National Inter-University Consortium for Telecommunications (CNIT), Bologna, 401 33 Italy
[5]Ericsson Research, Ericsson Telecomunicações Ltda., Indaiatuba, 13337-300 Brazil

CORRESPONDING AUTHOR: André L. P. Fernandes (e-mail: andrelpf@ufpa.br).

This work was supported by the Innovation Center, Ericsson Telecomunicações Ltda., Brazil, by the National Council for Scientific and Technological Development (CNPq), by the Coordination of Superior Level Staff Improvement (CAPES), and partly by the ECO-eNET project funded from the Smart Networks and Services Joint Undertaking (SNS JU) under grant agreement No. 10113933.

ABSTRACT User-centric (UC) distributed massive multiple-input multiple-output (D-mMIMO), also known as cell-free mMIMO, is a pivotal technology for enabling future mobile communication systems. While UC D-mMIMO intrinsically follows a distributed architecture, its processing can be implemented in a distributed or centralized fashion. This paper proposes a comprehensive cost assessment methodology for UC D-mMIMO, capturing its total cost of ownership and factoring in the deployment configuration, processing implementation, computational demands, and fronthaul signaling. The methodology considers two transmission reception point (TRP) deployment strategies. The first focuses only on supporting user equipment (UE) demands, while the other fulfills these requirements and also actively strives to provide a fairer service among UEs. The proposed methodology is then used to perform a techno-economic assessment of the feasibility of centralized versus distributed processing functional splits while varying key costs and TRP capabilities, like antenna and served UE count. Results suggest that with the TRP deployment that only supports the required UE rate, distributed processing is usually the most feasible option for UE demands of up to 50 Mbps, and centralized processing is more cost-effective in other cases. Additionally, when considering the actively fairer TRP deployment, centralized processing becomes cheaper for any UE demands.

INDEX TERMS Cell-free massive MIMO, feasibility analysis, network deployment, functional splits, techno-economic assessment, total cost of ownership.

I. INTRODUCTION

User-centric (UC) distributed massive multiple-input multiple-output (D-mMIMO), commonly called cell-free massive multiple-input multiple-output (mMIMO), emerges as a promising technology to meet the evolving needs of future mobile communication systems, like sixth-generation (6G) [1], [2]. It employs a large number of transmission-reception points (TRPs) scattered across a coverage area, each equipped with one or more antennas. In such a way that distinct TRPs engage in coordinated communication with different users' equipment (UEs), collectively processing the UEs signals by exchanging information through fronthaul links. To this end, the system leverages one or more edge cloud central processing units (CPUs), which facilitate the information exchange, perform baseband functions, and orchestrate the system's overall coordination [3], [4].

The unique combination of distributed deployment and joint signal processing culminates in macro-diversity gain, fostering higher network densification while maintaining interference at controllable levels. Consequently, these features pave the way for superior and more uniform spectral efficiency (SE) across the coverage area, overshadowing the performance of co-located mMIMO [3], [4].

The architecture of UC D-mMIMO is inherently distributed. However, the processing implementation can be either centralized or distributed. This flexibility arises from performing certain baseband functions locally at the TRPs or at the edge CPU. The processing is distributed when tasks such as channel estimation and precoding computation occur at the TRPs. This approach employs simpler precoding techniques, aligning with the system's distributed nature and offering high computational resource efficiency. Conversely, centralized processing enables more advanced processing techniques by performing the aforementioned tasks at edge CPUs, potentially achieving superior performance at the expense of increased computational complexity [4], [5].

Initially, UC D-mMIMO was mainly based on distributed processing due to its simplicity, which was believed to increase the system's scalability and reduce fronthaul signaling [3]. However, it was later proven that centralized processing could also be scalable. Moreover, it has potentially lower fronthaul signaling than the distributed case while providing much higher performance [5], [6]. Nevertheless, this does not mean centralized approaches are always superior. The computational complexity can be orders of magnitude higher than the distributed case [7]. Besides that, the fronthaul requirements can be higher in the centralized case depending on the antenna count on the TRP, its supported number of UEs, and the adequate sample bit width for the supported UE data rate [5], [8].

A comprehensive techno-economic comparison is essential to adequately assess the superiority of centralized or distributed processing in different situations. This analysis should scale factors like deployment expenses and power consumption with the required computational complexity and fronthaul signaling load, quantifying costs to support different UE traffic demands. However, the field of UC D-mMIMO's techno-economics remains largely uncharted in the literature [9]–[11]. The main reason behind this fact is the novelty of UC D-mMIMO as a theoretical concept operating under a new communication paradigm. Nevertheless, recent advancements in models for UC D-mMIMO clarified its capabilities and requirements [6]–[8], [12], [13]. These developments pave the way for a comprehensive techno-economic analysis.

This paper proposes a cost assessment methodology for UC D-mMIMO. The aim is to compare centralized and distributed processing implementations to determine their general feasibility. It also identifies specific scenarios where feasibility trends might differ. Furthermore, the cost trends are assessed for TRPs with varying capabilities, specifically antenna count and UE support. To this end, existing literature models had to be adapted and integrated with newly developed models for the deployment of the UC D-mMIMO system and its associated components.

A. LITERATURE REVIEW

1) UC D-MMIMO

As mMIMO matured into the primary solution for enhancing SE for fifth-generation (5G) systems, the research focus has shifted towards coordinated transmission techniques under the name cell-free mMIMO [6]. This new transmission approach is effectively equivalent to a UC distributed multiple-input multiple-output (MIMO) system, utilizing various TRPs to serve different UEs while still being rooted in technologies initially developed for traditional cellular mMIMO [3]. The UC communication ensures that UEs are in communication with a dynamically tailored subset of TRPs based on their individual needs. This approach eliminates fixed associations between UEs, TRPs, and coverage areas, virtually eliminating cell boundaries [3], [6].

In [14], a fully distributed, scalable UC architecture for D-mMIMO systems was introduced. The study advocated for distributed strategies in signal processing and power control, driven by the belief that the natural distributed architecture of UC D-mMIMO can deliver excellent performance using simple conjugate beamforming precoders, which are inherently scalable. The study also pointed out that centralized processing strategies may offer superior performance. However, they were deemed unnecessary for UC D-mMIMO, being unscalable and potentially burdensome on the fronthaul signaling.

Contrary to these beliefs, [5] shattered the notion that D-mMIMO consistently outperforms small-cell systems when relying solely on distributed conjugate beamforming. Moreover, the study also identified that centralized processing could potentially have a lower fronthaul signaling load than its distributed counterparts. Due to these characteristics, the work advocated for centralized processing, pointing out that local minimum mean square error (MMSE) precoders should be considered if distributed processing is pursued, as they consistently outperform small-cell systems. However, it is essential to note that the work did not consider scalability aspects. Additionally, it recognized that a non-infinite precision fronthaul, with an adequate representation of the sample bit width, can potentially alter the fronthaul bit rate behavior.

In [8], the implications of quantized signals on fronthaul of D-mMIMO networks across uplink and downlink were examined. This study modeled quantization-related errors using an additive quantization noise model (AQNM) based on Bussgang decomposition, presenting models for two functional splits representing distributed and centralized processing implementations. These models aligned with those in [5], accommodating a variable bit width depending on the number of UEs and fronthaul capacity. The results

corroborated with [5], proving that fronthaul signaling was smaller in the centralized implementation for a similar level of UE rate performance.

In [6], a scalable framework for UC D-mMIMO was introduced, containing modified centralized precoders, UC TRPs cluster formation, and pilot assignment. The study proved that centralized processing can be scalable, cementing its position as the best processing approach. Expanding on the framework, [7] identified that an increasing number of TRPs might reintroduce non-scalability. Accordingly, it complemented the TRPs cluster formation to guarantee scalability under such conditions.

Even with centralized approaches being appointed as the best ones, the distributed local partial MMSE (LP-MMSE) implementation is still being investigated as UC D-mMIMO systems delve into practical aspects, as it has less computational complexity and more flexibility of implementation [12], [13], [15].

2) 5G AND UC D-MMIMO TECHNO-ECONOMICS

In [16], an extensive analysis of literature concerning 5G techno-economics was carried out. This review considers various technologies, use cases, and evaluation metrics. The study's primary aim was to provide recommendations for techno-economic assessments of next-generation mobile communication systems. Several essential conclusions were reached. Firstly, the accuracy and reliability of any techno-economic analysis hinge on a well-defined network dimensioning procedure. Secondly, when evaluating financial metrics, it is imperative to consider both capital expenditure (CAPEX) and operational expenditure (OPEX). Lastly, a sensitivity analysis is essential to ascertain the validity of proposed models and methods. An important observation is that most of the works in the review did not focus on mMIMO, despite it being an integral enabler of the 5G.

Simplified techno-economic models on UC D-mMIMO systems were recently introduced by [9] and [10]. In [9], the feasibility of using serial interconnection among TRPs was evaluated, a possible solution to reduce the number of fronthaul links and decrease network complexity. The analysis focused on a fiber-based transport network, and the results suggested that a serial interconnection can be cost-effective in a tree configuration with two or three serially connected TRPs. However, the study points out that serial interconnection may not be feasible in high-demand scenarios. Limitations of this work include excessive emphasis on transport infrastructure and simplified models. For instance, only a basic conjugate beamforming distributed processing approach is considered, and no relationship between computational requirements and costs is delineated.

In [10], a comparative analysis of cost efficiency was presented, evaluating UC D-mMIMO against small cells. The study investigates various sizes of TRP clusters for each UE and examines different fiber transport connections under single and multiple CPU scenarios. The findings suggested that UC D-mMIMO can 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 has 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. In the latter case, the calculations scale solely with the size of the subset of TRPs serving each UE.

One of the primary limitations of [10] is addressed by [11], which expanded the analysis to incorporate energy-related OPEX in the model. Nevertheless, this subsequent work did not address the other deficiencies in the initial model. Furthermore, different types of OPEX costs still need to be explored. Although the energy model can be considered adequate, there is room for expansion, as many computational operations at the CPU and TRP are overlooked.

Finally, neither [9] nor [11] address the dimensioning of the necessary number of TRPs concerning demands. Instead, these works circumvent this challenge by assuming a fixed number of TRPs and delving into other aspects, like transport network configuration. However, it is evident that existing literature's dimensioning procedures for cellular systems, like the ones in [17] and [18], are 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 UC D-mMIMO can be imperative for future works.

B. CONTRIBUTIONS

The contributions of this paper can be summarized as:

- A cost assessment methodology is proposed to calculate the total cost of ownership (TCO) of UC D-mMIMO networks. One that scales components deployment expenses and power consumption with the required computational complexity and fronthaul signaling load. The cost results depend on the scenario, active UE load profile, and target UE expected rate.
- A comprehensive cost model is presented, covering both CAPEX and OPEX considerations. For CAPEX, expenses consider the acquisition and installation of (i) TRPs, (ii) edge cloud CPU, and (iii) fronthaul equipment. On the OPEX side, expenses take into account (i) repairs, (ii) equipment occupied floor space rent, and (iii) power consumption. Besides that, technician salaries impact both CAPEX and OPEX.
- A TRP deployment model is proposed to determine the necessary number of active TRPs based on coverage or capacity constraints. In the second case, the model supports a given expected UE rate derived from UE average rates or a proportional fairness-based UE rate. This latter metric complies with a service level agree-

ment (SLA), aiming to maintain a significant part of an agreed rate throughout a large portion of the coverage area.
- The number of baseband processing operations is adequately allocated between CPU and TRPs for two commonly adopted functional splits for UC D-mMIMO, which aim to support distributed and centralized processing implementations.
- A fronthaul bit rate calculation based on a maximum acceptable SE degradation due to quantization is proposed. Under this novel metric, two sub-optimal methods for bit allocation in distributed and centralized processing implementations are presented. A non-limited fronthaul bit rate for a negligible degradation in UE experience can be obtained using these methods.
- A non-vendor-specific model for the structure of TRPs is proposed and is used in conjunction with a cloud radio access network (C-RAN) workload consolidation model. This approach allows the derivation of equations for deployment expenses and energy consumption associated with TRPs and CPUs, while considering processing and fronthaul requirements.
- A techno-economic analysis is conducted in a dense urban scenario to compare distributed and centralized processing implementation for the downlink operation of UC D-mMIMO. The results show that the distributed method might offer cost benefits for demands up to 50 Mbps per UE, or even 200 Mbps when the TRP features at least seven antennas. Despite this, the centralized approach often presents greater cost-efficiency, especially in high-demand scenarios and when an actively fairer TRP deployment is utilized.
- Tree precoders are analyzed in terms of cost: LP-MMSE, partial regularized zero-forcing (P-RZF), and partial MMSE (P-MMSE). The first is implemented in a distributed fashion, and the two later in a centralized fashion. The results show that, in most cases, P-MMSE is more cost-efficient. However, P-RZF is the most feasible under high demands, with 500 Mbps per UE and a TRP antenna count larger than four.

C. PAPER OUTLINE AND NOTATIONS

The remainder of this paper is organized as follows. Section II presents the system model, detailing channel modeling, fronthaul constraints impact, channels estimation procedure, system scalability considerations, and UE rate calculation. Section III discusses the cost assessment methodology, modeling the required TRP count, fronthaul bit rate, and computational resource requirements. Moreover, the price and energy consumption models for TRPs and the CPU are also presented. Section IV introduces the cost models utilized to determine the TCO of the UC D-mMIMO system in the proposed methodology. Section V presents the results of this work for a baseline scenario and relevant variations in the assumptions. Finally, Section VI concludes the paper.

Notation: Boldface lowercase and uppercase letters denote vectors and matrices, respectively, the superscript $(\cdot)^{\mathrm{H}}$ denotes the conjugate-transpose operation, the $N \times N$ identity matrix is $\mathbf{I}_N$, and 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, and the notation $\mathcal{CN}(\mu, \sigma^2)$ stands for a complex Gaussian random variable with mean μ and variance σ^2.

II. SYSTEM MODELS

It is considered a downlink (DL) transmission of a UC D-mMIMO system with L TRPs with N antennas serving K single-antenna spatially distributed UEs. The TRPs are connected to an edge cloud CPUs via dedicated fronthaul links, and UEs are only served by best possible set of TRPs, as shown in Fig. 1. The system operates under time-division duplex (TDD) protocol inside a coherence time-frequency resource block with τ_c samples [4]. Moreover, details for channel modeling, fronthaul constraints impact, channel estimation procedure, system scalability considerations, and UE rate calculation are presented in the following subsections. For the reader's convenience, Table 1 lists all the mathematical representations used throughout the equations of this section.

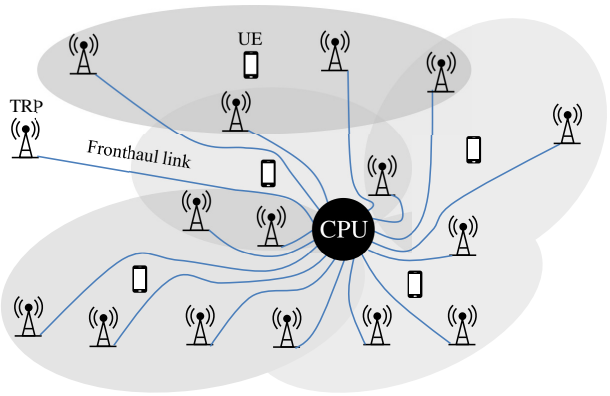


FIGURE 1. Illustration of the system model considered network architecture. Dedicated fronthaul links connect the edge cloud CPU to the TRPs. UEs are served by a limited optimal set of TRPs with available resources.

1) CHANNEL MODEL

The channel between the TRP l and the UE k ($\mathbf{h}_{l,k} \in \mathbb{C}^{N\times 1}$) undergoes independent correlated Rician fading in each coherence block, being defined as

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

where $\overline{\mathbf{h}}_{l,k} \in \mathbb{C}^{N\times 1}$ represents the LoS component, and $\mathbf{g}_{l,k} \sim \mathcal{CN}(\mathbf{0}_N, \mathbf{R}_{l,k}) \in \mathbb{C}^{N\times 1}$ denotes the NLoS component. The covariance matrix $\mathbf{R}_{l,k} = \mathbb{E}\{\mathbf{g}_{l,k}\mathbf{g}_{l,k}^{\mathrm{H}}\} \in \mathbb{C}^{N\times N}$

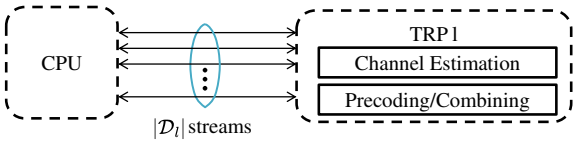
TABLE 1. List of mathematical notations used in Section II.

Symbol	Description
$\alpha_{l,k}$	Fronthaul quantization distortion factor between TRP l and UE k
α_l	Fronthaul quantization distortion factor in TRP l
$\beta_{l,k}$	Average channel gain between TRP l and UE k
$d_{l,k}$	Distance between TRP l and UE k
DS_k	Desired signal component for UE k
$\mathbf{g}_{l,k}$	Non-line-of-sight (NLoS) Ricean component
$\mathbf{h}_k$	Global channel of UE k
$\mathbf{h}_{l,k}$	Channel between TRP l and UE k
$\mathbf{h}_{l,k}^{\text{LoS}}$	Line-of-sight (LoS) channel between TRP l and UE k
$\mathbf{h}_{l,k}^{\text{NLoS}}$	NLoS channel between TRP l and UE k
$\overline{\mathbf{h}}_{l,k}$	LoS Ricean component
IS_k	Interferent signal component for UE k
K	Number of UEs
$\kappa_{l,k}$	Rician factor between TRP l and UE k
L	Number of TRPs
$\mathcal{M}_k$	Set of TRPs serving UE k
N	Number of antennas in each TRP
$\mathbf{q}_l$	Additive quantization noise in antenna signals for TRP l in a baseband processing at the CPU (BCPU)
$q_{l,i}$	Additive quantization noise for the signal of UE i on TRP l in a baseband processing at the TRP (BTRP)
p_{LoS}	LoS probability determined by the propagation scenario
$\mathbf{R}_{l,k}$	Covariance matrix with channel components and spatial correlation
$\rho_{l,k}$	DL power allocated by TRP l to UE k
σ_{dl}^2	DL additive white Gaussian noise (AWGN) noise
τ_c	Number of samples in the coherence block
τ_p	Number of orthogonal pilots
$\mathcal{D}_l$	Set of UEs served by TRP l
$\mathbf{D}_{l,k}$	Binary diagonal matrix indicating which antennas of TRP l serve UE k
QN_k	Fronthaul quantization noise component for UE k
$\overline{\mathbf{w}}_{l,k}$	Normalized precoder of UE k in TRP l
$\overline{\mathbf{w}}_k$	Normalized global precoder for UE k

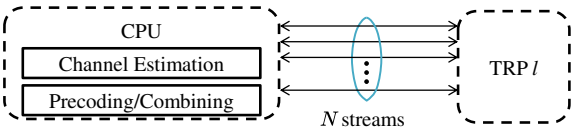
describes the spatial correlation [19]. Moreover, $\kappa_{l,k}$ is the Rician factor, which is modeled as a function of the distance $d_{l,k}$ between TRP l and UE k. It takes the minimum value between $10^{1.3-0.003d_{lk}}$ and $p_{\text{LoS}}(d_{l,k})/(1-p_{\text{LoS}}(d_{l,k}))$, where p_{LoS} is the LoS probability determined by the propagation scenario. Besides that, $\beta_{l,k}$ represents the average channel gain of $\mathbf{h}_{l,k}$, encompassing path loss and shadowing [19].

2) FRONTHAUL CONSTRAINTS IMPACTS

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 errors associated with the quantization processes are obtained from an AQNM applied to two different functional split approaches between CPU and TRP, presented in Fig. 2 [8]. In the first case, BTRP, the data symbols to each UE the TRP serves are quantized and sent by the CPU. Moreover, the signal to each antenna comes from the precoding procedure made at the TRPs. In this way, the number of data streams on the fronthaul of a TRP l is directly equivalent to the size of the set of UEs served by the said TRP, represented by $\mathcal{D}_l \subset \{1, ..., K\}$. In the second case, BCPU, the signals transmitted in each TRP antenna are quantized and sent by the CPU, which performs channel estimation and precoding. In this way, the number of data streams on the fronthaul equals N. Throughout this work, centralized processing implementations use the BCPU approach, and distributed processing implementations use the BTRP approach.



(a) BTRP



(b) BCPU

FIGURE 2. Simplified overview of the classical functional splits for UC D-mMIMO [8].

Beyond its impact on data transmission, the fronthaul's capacity limitations also extend to the channel estimation procedure. Specifically, under the BCPU approach, the pilot samples arriving at the CPU undergo distortion due to quantization during fronthaul transmission. This introduces an additional source of error to the channel estimation process, subsequently leading to compromised precoder performance and reduced SE [8].

3) CHANNEL ESTIMATION PROCEDURE

The TRP uplink (UL) channel estimates to its UEs are obtained through an MMSE estimator using UL orthogonal pilots transmitted by the UEs, while accounting for quantization distortion on the pilot samples in centralized processing approaches [8]. DL channel estimation is not performed since the channel reciprocity of using TDD inside a coherence block allows UL channel estimates to be used for DL processing [3]. The number of orthogonal pilots equals the number of samples in each pilot (τ_p), which theoretically can be as large as τ_c. Still, in practice, τ_p is smaller to avoid over-signaling in the coherence block and due to hardware limitations in the devices performing the channel estimation procedure. If the total number of UEs is larger than τ_p, pilot contamination happens due to pilot reuse among UEs. This

contamination degrades the estimation quality and generates DL coherent interference [3], [6].

4) SYSTEM SCALABILITY

A dynamic cooperation clustering (DCC) framework is used to manage the UE/TRP connections and assure that the computational complexity of channel estimation and signal processing will be limited even if L or K grows to infinity. The initial access procedure from [6] is executed combined with the TRP cluster size control technique outlined in [7]. This strategy ensures that each TRP establishes connections with a maximum of τ_p UEs, and correspondingly, each UE forms connections with up to $U_{\max}$ TRPs. Moreover, scalable LP-MMSE, P-RZF, and P-MMSE precoders are utilized [4].

For mathematical representation, a diagonal matrix $\mathbf{D}_{l,k}$ is used to indicate which antennas of a TRP l serve UE k. Besides that, it is considered that all antennas of a TRP will provide connection to all its served UEs [6]. In this way, $\mathbf{D}_{l,k}$ can be expressed as follows

$$\mathbf{D}_{l,k} = \begin{cases} \mathbf{I}_N, \text{ if } (l) \in \mathcal{M}_k \\ \mathbf{0}_N, \text{ otherwise} \end{cases}, \tag{2}$$

where $\mathcal{M}_k \subset \{1, ..., L\}$ represents the set of TRPs connected to each UE k. $\mathcal{M}_k$ is complemented by $\mathcal{D}_l$.

5) UE RATE CALCULATION

A lower bound of the DL SE can be obtained using the use-and-then-forget bound under decoders based on channel hardening while considering the fronthaul quantization noise and distortion [6], [8], i.e.,

$$\mathrm{SE}_k = \left(1 - \frac{\tau_p}{\tau_c}\right) \log_2 \left(1 + \frac{\mathrm{DS}_k}{\mathrm{IS}_k - \mathrm{DS}_k + \mathrm{QN}_k + \sigma^2_{\mathrm{dl}}}\right), \tag{3}$$

where DS_k, IS_k, QN_k and σ^2_{dl} are the powers of the desired signal, interference signals, fronthaul quantization noise, and AWGN noise, respectively. The values for the first three variables are calculated as

$$\begin{aligned} \mathrm{DS}_k &= \left| \sum_{l=1}^{L} \alpha_{l,k} \mathbb{E}\left\{ \sqrt{\rho_{l,k}}\, \mathbf{h}^{\mathrm{H}}_{l,k} \mathbf{D}_{l,k} \overline{\mathbf{w}}_{l,k} \right\} \right|^2, \\ \mathrm{IS}_k &= \sum_{i=1}^{K} \mathbb{E}\left\{ \left| \sum_{l=1}^{L} \alpha_{l,k} \sqrt{\rho_{l,i}}\, \mathbf{h}^{\mathrm{H}}_{l,k} \mathbf{D}_{l,i} \overline{\mathbf{w}}_{l,i} \right|^2 \right\}, \\ \mathrm{QN}_k &= \begin{cases} \mathbb{E}\left\{ \left| \sum_{l=1}^{L} \mathbf{h}^{\mathrm{H}}_{l,k} \mathbf{D}_{l,i} \mathbf{q}_l \right|^2 \right\}, \text{ for BCPU} \\ \mathbb{E}\left\{ \left| \sum_{l=1}^{L} \mathbf{h}^{\mathrm{H}}_{l,k} \mathbf{D}_{l,i} \sum_{i=1}^{K} \overline{\mathbf{w}}_{l,i} q_{l,i} \right|^2 \right\}, \text{ for BTRP} \end{cases} \end{aligned}, \tag{4}$$

where $\alpha_{l,k}$ is the quantization distortion factor between TRP l and UE k. In the context of a BCPU implementation, $\alpha_{l,k} = \alpha_l$, indicating that distortion occurs only at the TRP level. Additionally, $\rho_{l,k}$ is the DL power allocated by TRP l to UE k. Moreover, $\overline{\mathbf{w}}_{l,k} \in \mathbb{C}^{N \times 1}$ represents the unit-power precoding vector for the channel between TRP l and UE k. Centralized precoders like P-MMSE and P-RZF usually are calculated for the collective channel $\mathbf{h}_k = [\mathbf{h}^{\mathrm{T}}_{1,k}, \cdots, \mathbf{h}^{\mathrm{T}}_{L,k}]^{\mathrm{T}}$, resulting in global UE precoder $\overline{\mathbf{w}}_k \in \mathbb{C}^{LN \times 1}$ [4]. Despite this, the individual $\overline{\mathbf{w}}_{l,k}$ can still be obtained from $\overline{\mathbf{w}}_k$ since $\overline{\mathbf{w}}_k = [\overline{\mathbf{w}}^{\mathrm{T}}_{1,k}, \cdots, \overline{\mathbf{w}}^{\mathrm{T}}_{L,k}]^{\mathrm{T}}$. Finally, $\mathbf{q}_l \sim \mathcal{CN}\big(0, \alpha_l (\alpha_l - 1) \sum_{k=1}^{K} \rho_{l,k} \mathbb{E}\{\overline{\mathbf{w}}_{l,k} \overline{\mathbf{w}}^{\mathrm{H}}_{l,k}\}\big)$ denotes the additive quantization noise in the antenna signals for TRP l in a BCPU implementation, and $q_{l,i} \sim \mathcal{CN}\left(0, \alpha_{l,i} (\alpha_{l,i} - 1) \rho_{l,i}\right)$ represents the additive quantization noise for the signal of UE i on TRP l in a BTRP implementation.

III. COST ASSESSMENT METHODOLOGY

The proposed methodology to assess the total cost of a UC D-mMIMO is presented in Fig. 3. 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.

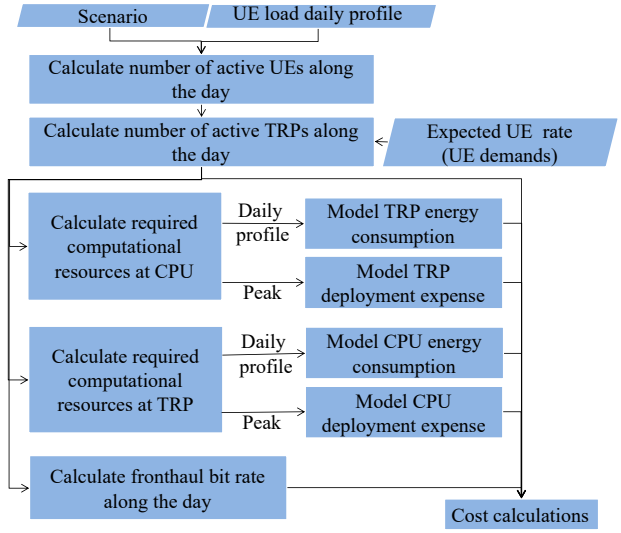


FIGURE 3. Proposed cost assessment methodology of a UC D-mMIMO system.

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 UC D-mMIMO system.

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.

A. NUMBER OF ACTIVE TRPS

This subsection calculates the number of required TRPs to support the UE's requirements. For the reader's convenience, Table 2 outlines all mathematical notations introduced by equations throughout the subsection.

TABLE 2. List of mathematical notations introduced in Subsection III-A.

Symbol	Description
α_t	Active UE load ratio at time t
F_{cov}	Percentage of the SLA agreed coverage area with at least the guaranteed rate
F_{rate}	Percentage of the SLA agreed rate equivalent to the network rate guarantee
L_{max}	Maximum number of possible deployed TRPs
L_t	Number of active TRPs to support the UEs load at time t
$L_{t,C}$	Minimum number TRPs to support all UEs inside the coverage area at time t
$L_{t,R}$	Required number of TRPs to deliver a UE expected rate R at time t
$\mathcal{L}$	Specific set of TRPs counts generating $\mathcal{R}$
R	UE expected rate inside the coverage area
R_{agreed}	SLA agreed rate
R_{acov}	Average of the achievable UE rates higher or equal to $r_{F_{\text{cov}}}$
R_{bcov}	Average of the achievable UE rates smaller than $r_{F_{\text{cov}}}$
$\mathcal{R}$	Set of rates R for a specific set of TRP counts
$r_{F_{\text{cov}}}$	$(100 - F_{\text{cov}})$ th percentile rate in the UE achievable rate cumulative distribution function (CDF)
ρ_K	UE density
S	Scenario area

A scalable UC D-mMIMO system with DCC ensures that each UE remains connected to at least one TRP [4]. 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 work, this capacity corresponds to the number of pilots [6]. However, to effectively enhance the capacity for UEs in a UC D-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 [3].

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 [20] as

$$L_t = \max\left(L_{t,C}, L_{t,R}\right), \tag{5}$$

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

Assuming that each individual TRP can have a effective communication channel to any UE in the entire coverage area, then $L_{t,C} = \rho_K \alpha_t \, S/\tau_p$, 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 $L_{t,C}$ is more complex and not considered in this work, being left for future implementations[1].

There is no straightforward way to compute $L_{t,R}$. Nevertheless, obtaining an average rate equivalent to R for a given $L = L_{t,R}$ and $K = \rho_K \, \alpha_t S$ is relatively simple using a Monte Carlo simulation process in conjunction with (4) [4]. In this context, it is possible to calculate a set of rates $\mathcal{R}$ for a specific set of TRP counts, defined by $\mathcal{L} = \{L_{t,C}, L_{t,C} + L_{\text{step}}, L_{t,C} + 2L_{\text{step}}, \cdots, L_{\max}\}$, where $L_{\max}$ is the maximum value of TRPs that can be implemented and L_{step} is the increment step for each element in $\mathcal{L}$. This procedure results in $\mathcal{R} = \{R_1, R_2, \cdots, R_{|\mathcal{L}|}\}$ where $R_1 < R_2 < \cdots < R_{|\mathcal{L}|}$. Finally, the value for an arbitrary $L_{t,R}$ can be calculated using an interpolation process, which takes $\mathcal{L}$ and $\mathcal{R}$ as inputs, as long as $R_1 < R < R_{|\mathcal{L}|}$.

A rate R based on the average UE rate 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 SLA with an agreed UE rate [21]. Ensuring a fixed rate in mobile networks is challenging due to UEs' mobility and other random factors [22]. 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 [4], the agreed rate is calculated by

$$R_{\text{agreed}} = \min\left(R_{\text{acov}}, \frac{r_{F_{\text{cov}}}}{0.01 F_{\text{rate}}}\right), \tag{6}$$

[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 $L_{t,C}$ is defined as the number of clusters.

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

$$R = \frac{100 - F_{\text{cov}}}{100} R_{\text{bcov}} + \frac{F_{\text{cov}}}{100} R_{\text{agreed}}, \tag{7}$$

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.

B. FRONTHAUL BIT RATE CALCULATION

This subsection computes the necessary fronthaul data rate for each TRP to meet the UE's demands. To aid the reader, Table 3 summarizes all mathematical symbols introduced in the equations within this subsection.

TABLE 3. List of mathematical notations introduced in Subsection III-B.

Symbol	Description
a_{deg}	Maximum acceptable SE degradation due to quantized fronthaul samples in bps/Hz
B	System bandwidth
b_l^{data}	Fronthaul bit width for the data symbols in all antennas of TRP l for the BCPU implementation
$b_{l,k}^{\text{data}}$	Fronthaul bit width for the data symbols between TRP l and UE k in the BTRP implementation
b_l^{pil}	Fronthaul bit width for pilot samples for channel estimation in all antennas of TRP l for the BCPU implementation
$F_{l,t}$	Required fronthaul bandwidth for TRP l at time t
$\mathcal{D}_{l,t}$	Set of UEs served by TRP l at time t

In the context of UC D-mMIMO, the calculation of fronthaul bit rate relies on multiple factors. These include 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 [4], [6], [8].

In this work, the transmission of statistics is disregarded since the interval for changes in the channel statistics is usually much larger than the coherence time [3]. The fronthaul bit rate of the two considered split implementations for the UE load of the time t is given by

$$F_{l,t} = \begin{cases} 2B\left(1 - \frac{\tau_p}{\tau_c}\right) \sum\limits_{k \in \mathcal{D}_{l,t}} b_{l,k}^{\text{data}}, & \text{for BTRP} \\ 2NB\left[\left(1 - \frac{\tau_p}{\tau_c}\right) b_l^{\text{data}} + \frac{\tau_p}{\tau_c} b_l^{\text{pil}}\right], & \text{for BCPU} \end{cases} \tag{8}$$

where B is the total available bandwidth, $\mathcal{D}_{l,t}$ is $\mathcal{D}_l$ at time t, and $b_{l,k}^{\text{data}}$ is the bit width for the data symbols inside the coherence block between TRP l and UE k in the BTRP implementation [4], [8]. Moreover, for the BCPU implementation, b_l^{data} is the bit width for the data samples of the coherence block in all antennas of TRP l, and b_l^{pil} is the bit width of pilot samples for channel estimation. The latter is applied only to τ_p samples of the coherence block [4], [8]. 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 [5].

In the literature, the bit width is usually pre-fixed or calculated for a given fronthaul capacity [8]. The cost-analysis nature of this work allows 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 work 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 $b_l^{\text{data}} = b_{l'}^{\text{data}} \forall l' \in \{1, ..., L\}$, $b_l^{\text{pil}} = b_{l'}^{\text{pil}} \forall l' \in \{1, ..., L\}$ and $b_{l,k}^{\text{data}} = b_{l',k}^{\text{data}} \forall l' \in \{1, ..., L\}$, the Algorithms 1 and 2 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.

C. REQUIRED COMPUTATIONAL COMPLEXITY CAPACITY IN CPU and TRPS

This subsection computes the necessary capacity in terms of giga operations per second (GOPS) that the hardware of CPU and TRPs will require to operate in the BCPU and BTRP splits. To facilitate understanding, Table 4 provides a summary of all mathematical symbols introduced throughout the subsection.

Depending on the functional split, various digital signal processing procedures are executed at the TRP or the CPU, as shown in Fig. 4, which presents the task division for two commonly adopted functional splits for UC D-mMIMO [8]. Consequently, the computational complexity of tasks performed at the TRP or CPU varies according to the chosen functional split [8], [13].

In both BCPU and BTRP cases, certain operations are always executed at the CPU, and the number GOPS associated with these operations is calculated using a reference scaling model [23]. Table 5 provides a detailed breakdown of the scaling factors used in these calculations. Within this context, the variables B_{base} and SE_{base} represent the bandwidth and SE of the reference GOPS value. In contrast, B and $\text{SE}_{t,R}$ represent the adopted bandwidth and SE, with the latter assumed to be equal to the simulated average SE for an expected UE rate R at the UE load of the time t. Final

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

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

Algorithm 2: Bit Allocation Evaluation in a BCPU split when the same bit width is applied at a TRP level ($b_{l}^{\text{data}} = b_{l'}^{\text{data}} \forall l' \in \{1, ..., L\}$).

Input: K, a_{deg}
$b_l^{\text{data}} \leftarrow 1$
$\textbf{SE} \leftarrow \text{CALCULATE\_SE}(b_l^{\text{data}})$ ▷ Calculates the SE of each UE according to the bit width
$\textbf{SE\_target} \leftarrow \text{CALCULATE\_SE}(\infty)$ ▷ Calculates the SE of each UE for a fronthaul with unlimited capacity
while $\max_{i=1,\cdots,K}(\textbf{\textit{SE\_target}}[i] - \textbf{\textit{SE}}[i]) > a_{\text{deg}}$ **do**
  $b_l^{\text{data}} \leftarrow b_l^{\text{data}} + 1$
  $\textbf{SE} \leftarrow \text{CALCULATE\_SE}(b_l^{\text{data}})$
end
Output: b_l^{data}

GOPS values are obtained by multiplying the scaling factor with their respective GOPS reference value.

The calculations in Table 5 follow the methodology in [23], with adjustments made to account for specific characteristics of UC D-mMIMO. In these systems, all UEs transmit/receive information using the entire bandwidth. In this way, the number of streams is equivalent to the number of UEs. Additionally, the total number of antennas equals LN, and the OFDM modulation scales with the number of UEs, unlike the reference, where it scales with the number of antennas. This deviation arises from the fact that in UC D-mMIMO, there is no need to modulate for each antenna since the precoding process takes each UE's modulated symbol as input.

TABLE 4. List of mathematical notations introduced in Subsection III-C.

Symbol	Description
B_{base}	Bandwidth of the reference GOPS values
$CC_{\text{all},t}^{\text{comb}}$	Number of complex multiplications and divisions in the CPU to generate the precoding vectors at time t
$CC_{l,t}^{\text{comb}}$	Number of complex multiplications and divisions in the TRP l to perform channel estimation at time t
$CC_{\text{all},t}^{\text{est}}$	Number of complex multiplications and divisions in the CPU to perform channel estimation at time t
$CC_{l,t}^{\text{est}}$	Number of complex multiplications and divisions in the TRP l to generate the precoding vectors at time t
f_s	Sampling frequency
γ_{Chcd}	GOPS scaling parameter for channel coding
γ_{HLct}	GOPS scaling parameter for higher-layer control
γ_{HLnt}	GOPS scaling parameter for higher-layer network
γ_{MpDp}	GOPS scaling parameter for layer mapping and demapping
γ_{OFDM}	GOPS scaling parameter for orthogonal frequency-division multiplexing (OFDM) modulation and demodulation
$\text{GOPS}_{\text{Chcd}}$	Reference GOPS value for channel coding
$\text{GOPS}_{\text{HLct}}$	Reference GOPS value for higher-layer control
$\text{GOPS}_{\text{HLnt}}$	Reference GOPS value for higher-layer network
$\text{GOPS}_{\text{MpDp}}$	Reference GOPS value for layer mapping
$\text{GOPS}_{\text{OFDM}}$	Reference GOPS value for OFDM modulation and demodulation
$\text{GOPS}_{t,R}^{\text{BCPU}}$	GOPS for specific CPU operations of the BCPU split for an expected UE rate R at time t
$\text{GOPS}_{t,R}^{\text{CPU}}$	GOPS to be executed at the CPU for an expected UE rate R at time t
$\text{GOPS}_{t,R}^{\text{CPUcommon}}$	GOPS to be executed at the CPU in both BCPU and BTRP splits for expected UE rate R at time t
$\text{GOPS}_{t,R}^{\text{TRP}}$	Required TRP processing capacity in GOPS to efficiently handle the UE rate R at time t
N_{DFT}	Dimension of the discrete Fourier transform (DFT)
N_{sc}	Number of subcarriers
$\mathcal{Q}_l$	Subset of UEs with TRPs in common with those served by TRP l
$\mathcal{S}_k$	Subset of UEs that are partially served by the same TRPs as UE k
SE_{base}	SE of the reference GOPS values
$\text{SE}_{t,R}$	Expected SE for an expected UE rate R at time t
T_s	OFDM symbol duration
$\mathcal{Z}_k$	Subset of TRPs serving the UEs that are in $\mathcal{S}_k$

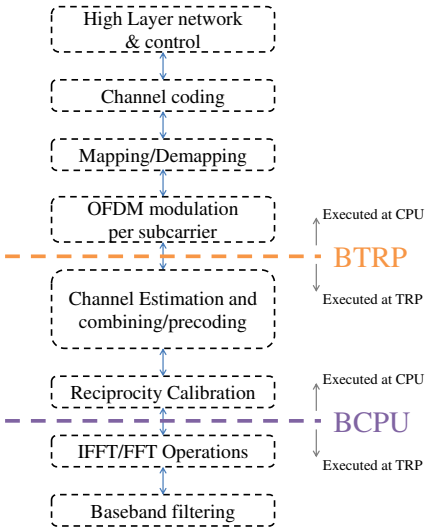


FIGURE 4. Distribution of digital signal processing procedures in the CPU and TRPs for BCPU and BTRP splits.

TABLE 5. GOPS scaling parameters calculation for common CPU operations in BCPU and BTRP [23].

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

The total summed GOPS associated with higher-layer control/network, channel coding, mapping/demapping, and OFDM modulation/demodulation for an expected UE rate R at the UE load of the time t are aggregated in $\text{GOPS}_{t,R}^{\text{CPUcommon}} = \gamma_{\text{HLnt}}\,\text{GOPS}_{\text{HLnt}} + \gamma_{\text{HLct}}\,\text{GOPS}_{\text{HLct}} + \gamma_{\text{Chcd}}\,\text{GOPS}_{\text{Chcd}} + \gamma_{\text{MpDp}}\,\text{GOPS}_{\text{MpDp}} + \gamma_{\text{OFDM}}\,\text{GOPS}_{\text{OFDM}}$, where $\text{GOPS}_{\text{HLnt}}$, $\text{GOPS}_{\text{HLct}}$, $\text{GOPS}_{\text{Chcd}}$, $\text{GOPS}_{\text{MpDp}}$, $\text{GOPS}_{\text{OFDM}}$ are the reference values of GOPS for higher-layer network, higher-layer control, channel coding, layer mapping and demapping, and OFDM modulation and demodulation, respectively. In this way, the 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

$$\text{GOPS}_{t,R}^{\text{CPU}} = \begin{cases} \text{GOPS}_{t,R}^{\text{CPUcommon}} + \text{GOPS}_{t,R}^{\text{BCPU}}, & \text{for BCPU} \\ \text{GOPS}_{t,R}^{\text{CPUcommon}}, & \text{for BTRP} \end{cases} \tag{9}$$

where $\text{GOPS}_{t,R}^{BCPU}$ 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

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

where N_{sc} is the number of subcarriers and T_s is the OFDM symbol duration. Moreover, $CC_{\text{all},t}^{\text{est}}$ and $CC_{\text{all},t}^{\text{comb}}$ 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. The term $8/(10^9 T_s)$ converts the number of complex multiplications to GOPS. Additionally, reciprocity calibration is a 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 $(\tau_c - \tau_p)/\tau_c$ [13].

Table 6 presents the values of $CC_{\text{all}}^{\text{est}}$ and $CC_{\text{all}}^{\text{comb}}$ for precoders considered in this work. The presented calculations are derived from [4]. In the table, $\mathcal{S}_k = \{i : \mathbf{D}_k\mathbf{D}_i \neq \mathbf{0}_{LN\times LN}\}$ represents the subset of UEs that are partially served by the same TRPs as UE k. 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 $\mathcal{Q}_l = \cup_{(l'\in\mathcal{M}_k)}\mathcal{D}_{l'}$ represents the UEs with TRPs in common with those served by TRP l. Both $\mathcal{Z}_k$ and $\mathcal{Q}_l$ are utilized to calculate common operations performed only once for each UE k or TRP l, such as channel estimation.

In both BCPU and BTRP cases, baseband filtering and inverse fast Fourier transform (IFFT)/fast Fourier transform (FFT) operations are executed at the TRP. The GOPS of these common operations for an expected UE rate R at the UE load of the time t can be calculated as

$$\text{GOPS}_{t,R}^{\text{TRPcommon}} = \underbrace{\frac{8N_{DFT}\log_2(N_{DFT})}{T_s 10^9}}_{\text{FFT/IFFT}} + \underbrace{\frac{40Nf_s}{10^9}}_{\text{Baseband Filter}}, \tag{11}$$

where N_{DFT} represents the dimension of the 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 [13].

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

$$\text{GOPS}_{t,R}^{\text{TRP}} = \begin{cases} \text{GOPS}_{t,R}^{\text{TRPcommon}}, & \text{for BCPU} \\ \text{GOPS}_{t,R}^{\text{TRPcommon}} + \text{GOPS}_{t,R}^{\text{BTRP}}, & \text{for BTRP} \end{cases} \tag{12}$$

where $\text{GOPS}_{t,R}^{\text{BTRP}}$ 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

$$\begin{aligned}\mathrm{GOPS}_{t,R}^{\mathrm{BTRP}} = &\underbrace{\frac{8N_{sc}CC_{l,t}^{\mathrm{est}}}{T_s 10^9 \tau_c}}_{\text{Channel estimation}} + \underbrace{\frac{8N_{sc}CC_{l,t}^{\mathrm{comb}}}{T_s 10^9 \tau_c}}_{\text{Combining computation}} \\ &+ \underbrace{\frac{8N_{sc}N\left|\mathcal{D}_{l,t}\right|}{T_s 10^9 \tau_c}}_{\text{Reciprocity calibration}} + \underbrace{\frac{8N_{sc}N\left(\tau_c-\tau_p\right)\left|\mathcal{D}_{l,t}\right|}{T_s 10^9 \tau_c}}_{\text{Precoding}},\end{aligned} \tag{13}$$

where $CC_{l,t}^{\mathrm{est}}$ and $CC_{l,t}^{\mathrm{comb}}$ denote the number of complex multiplications and divisions that the TRP l needs to perform channel estimation and generate the combining vectors for all active UEs at time t. Moreover, $CC_{l,t}^{\mathrm{est}}$ and $CC_{l,t}^{\mathrm{comb}}$ are computed as in Table 6.

D. TRP STRUCTURE MODEL

This subsection calculates the power consumption and expected prices for TRPs to accommodate the UE's requirements under both considered functional splits. For ease of reference, Table 7 outlines all mathematical notations utilized in equations within the subsection.

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. Fig. 5 provides an illustrative overview of the components of a non-vendor specific TRP in the UC D-mMIMO system. These components include antennas for bidirectional signal transmission, an analog front-end for initial radio signal processing, DSPs for tasks such as channel estimation and FFT/IFFTs conversions, and an I/Os interface that facilitates seamless network communication [3], [13], [24].

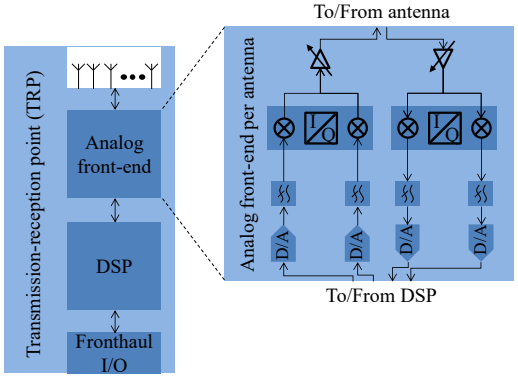


FIGURE 5. Example of the components for a non-vendor specific TRP in a UC D-mMIMO system.

The analog front end comprises several subcomponents, such as VGAs, IQ modulators, filters, DAC, and ADC converters. Fig. 5 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 SoC configuration [3], [13], [24].

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 [13], [23], [24]. In this context, it can be calculated as

$$P_{\mathrm{TRP}} = \mathrm{pw}_{\mathrm{DSP}} + \mathrm{pw}_{\mathrm{AFend}} + \mathrm{pw}_{\mathrm{IOint}} + \alpha_{\mathrm{amp}}\mathrm{pw}_{\mathrm{Tx}}, \tag{14}$$

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

$$\mathrm{pw}_{\mathrm{DSP}} = \gamma_{\mathrm{pwDcore}}\left\lceil \frac{\mathrm{GOPS}_{t,R}^{\mathrm{TRP}}}{\mathrm{CAP}_{\mathrm{Dcore}}} \right\rceil + \gamma_{\mathrm{pwDSP}}\,\mathrm{GOPS}_{t,R}^{\mathrm{TRP}} + \mathrm{pw}_{\mathrm{DSP}}^{\mathrm{other}}, \tag{15}$$

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. $\mathrm{GOPS}_{t,R}^{\mathrm{TRP}}$ denotes the required TRP 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, respectively. The analog front-end power consumption is given by

$$\mathrm{pw}_{\mathrm{AFend}} = 2N\left(2\mathrm{pw}_{\mathrm{filter}}^{\mathrm{ana}} + \mathrm{pw}_{\mathrm{IQmod}} + \mathrm{pw}_{\mathrm{VGA}} + \mathrm{pw}_{\mathrm{DAC}} + \mathrm{pw}_{\mathrm{ADC}}\right), \tag{16}$$

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

$$\mathrm{pr}_{\mathrm{TRP}} = \mathrm{pr}_{\mathrm{DSP}} + \mathrm{pr}_{\mathrm{AFend}} + \mathrm{pr}_{\mathrm{IOint}} + N\mathrm{pr}_{\mathrm{ant}}, \tag{17}$$

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

$$\mathrm{pr}_{\mathrm{DSP}} = \gamma_{\mathrm{prDcore}}\left\lceil \frac{\mathrm{GOPS}_{\mathrm{peak},R}^{\mathrm{TRP}}}{\mathrm{CAP}_{\mathrm{Dcore}}} \right\rceil + \mathrm{pr}_{\mathrm{DSP}}^{\mathrm{base}}, \tag{18}$$

where $\gamma_{\mathrm{prDcore}}$ is a price slope for the necessary number of cores in the DSP, $\mathrm{GOPS}_{\mathrm{peak},R}^{\mathrm{TRP}}$ is the peak number GOPS in a TRP to provide an expected UE rate R, and $\mathrm{pr}_{\mathrm{DSP}}^{\mathrm{base}}$ is a fixed price related to other DSP construction parameters. The analog front-end price is given by

$$\mathrm{pr}_{\mathrm{AFend}} = \alpha_{\mathrm{SoC}}2N(2\mathrm{pr}_{\mathrm{filter}}^{\mathrm{ana}} + \mathrm{pr}_{\mathrm{IQmod}} + \mathrm{pr}_{\mathrm{VGA}} + 2\mathrm{pr}_{\mathrm{DAC|ADC}}, \tag{19}$$

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.

TABLE 6. 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	$\sum_{l=1}^{L}\left(N\tau_p+N^2\right)\lvert\mathcal{Q}_l\rvert$	$\sum_{l=1}^{L}\left[\frac{1}{2}\left(\lvert\mathcal{Q}_l\rvert^2+\lvert\mathcal{Q}_l\rvert\right)N\right]+\sum_{k=1}^{K}\left[\lvert\mathcal{S}_k\rvert^2+N\lvert\mathcal{M}_k\rvert\lvert\mathcal{S}_k\rvert+\frac{1}{3}\left(\lvert\mathcal{S}_k\rvert^3-\lvert\mathcal{S}_k\rvert\right)+\lvert\mathcal{S}_k\rvert\right]$
P-MMSE	$\sum_{k=1}^{K}\left(N\tau_p+N^2\right)\lvert\mathcal{Z}_k\rvert$	$\sum_{k=1}^{K}\left[\frac{1}{2}\left(\left(N\lvert\mathcal{Z}_k\rvert\right)^2+N\lvert\mathcal{Z}_k\rvert\right)+\left(N\lvert\mathcal{M}_k\rvert\right)^2+\frac{1}{3}\left(\left(N\lvert\mathcal{M}_k\rvert\right)^3-N\lvert\mathcal{M}_k\rvert\right)+N\lvert\mathcal{M}_k\rvert\right]$
LP-MMSE	$\sum_{l=1}^{L}\left(N\tau_p+N^2\right)\lvert\mathcal{D}_l\rvert$	$\sum_{l=1}^{L}\left[\frac{1}{2}(N^2+N)\lvert\mathcal{D}_l\rvert+N^2\lvert\mathcal{D}_l\rvert+\frac{1}{3}(N^3-N)+N\right]$

TABLE 7. List of mathematical notations introduced in Subsection III-D.

Symbol	Description
α_{amp}	Expansion factor to account for losses in the amplification process
α_{SoC}	Price reduction factor due to System-on-a-Chip (SoC) integration
γ_{prDcore}	Price slope for the necessary number of cores in the digital signal processing (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
P_{TRP}	TRP power consumption
$\text{CAP}_{\text{Dcore}}$	GOPS capacity of the DSP processing core
$\text{GOPS}^{\text{TRP}}_{\text{peak},R}$	Peak number of GOPS in a TRP to provide an expected UE rate R
pr_{AFend}	TRP analog front-end price
pr_{ant}	Antenna price
$\text{pr}_{\text{DAC}\lvert\text{ADC}}$	Price of digital-to-analog converter (DAC) or analog-to-digital converter (ADC) in the TRPs
pr_{DSP}	DSP price
$\text{pr}^{\text{base}}_{\text{DSP}}$	Fixed price related to other DSP construction parameters
$\text{pr}^{\text{ana}}_{\text{filter}}$	TRP analog filter price
pr_{IOint}	I/O interface price
pr_{IQmod}	TRP in-phase and quadrature (IQ) modulator price
pr_{TRP}	TRP expected price
pr_{VGA}	TRP variable gain amplifier (VGA) price
pw_{ADC}	TRP ADC power consumption
pw_{AFend}	TRP analog front-end power consumption
pw_{DAC}	TRP DAC power consumption
pw_{DSP}	TRP DSP power consumption
$\text{pw}^{\text{other}}_{\text{DSP}}$	Non-GOPS dependant power consumption in the DSP
$\text{pw}^{\text{ana}}_{\text{filter}}$	TRP analog filter power consumption
pw_{IOint}	TRP input/output (I/O) interfaces power consumption
pw_{IQmod}	TRP IQ modulator power consumption
pw_{Tx}	TRP transmission power
pw_{VGA}	TRP VGA power consumption

E. CPU STRUCTURE MODEL

This subsection evaluates the power consumption and expected pricing for edge cloud CPU required to satisfy the UE requirements under the analyzed functional splits. For the convenience of the reader, Table 8 compiles all mathematical notations introduced in the equations within this subsection.

The CPU is deployed virtually in edge-cloud servers, following the C-RAN workload consolidation model outlined in [25]. The edge cloud CPU is then composed of global cloud controller (GCC), workload dispatcher, GPPs, and monitor/sensors, as presented in Fig. 6. 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

TABLE 8. List of mathematical notations introduced in Subsection III-E.

Symbol	Description
CAP_{bat}	CPU backup power battery capacity in Wh factoring in the depth of discharge
CAP_{GPP}	General purpose processor (GPP) processing capacity in GOPS
CAP_{rack}	Maximum amount of GPPs that an edge cloud CPU rack can hold
$\gamma_{\text{Co}\lvert\text{PD}}$	Price slope for CPU cooling and power distribution infrastructure
γ_{inv}	Price slope for inverter in the Edge CPU
$N^{\text{act}}_{\text{GPPs},t}$	Number of active GPPs in the edge cloud CPU at time t
N_{GPPs}	Number of GPPs deployed at the edge cloud CPU
$P_{\text{CPU},t}$	Edge cloud CPU power consumption at time t
$P^{\text{IT}}_{\text{CPU},t}$	Edge cloud CPU information technology (IT) equipment power consumption at time t
$P^{\text{cool}}_{\text{CPU},t}$	Edge cloud CPU cooling system power consumption at time t
$P_{\text{GPP},t}$	Power consumption of the GPP at a time t
$P^{\text{peak}}_{\text{CPU}}$	Edge cloud CPU peak power consumption
PUE_{cool}	Power usage effectiveness (PUE) of the edge cloud CPU cooling system
$\text{pr}^{\text{CPU}}_{\text{Sinf}}$	Price of the support infrastructure for the edge cloud CPU
pr_{bat}	Price for the battery's acquisition and installation in CPU deployment
$\text{pr}_{\text{rk\&nt}}$	Price for acquisition and installation cost of a rack and the network equipment in CPU deployment
$\text{pw}^{\text{idle}}_{\text{GPP}}$	Idle power consumption of the GPP
$\text{pw}^{\text{peak}}_{\text{GPP}}$	Peak power consumption of the GPP
$\text{pw}^{\text{rack}}_{\text{Net}}$	Power consumption of the network equipment per rack in CPU deployment
s_{CPU}	Deployment area of the edge cloud CPU
s_{rack}	Necessary area to install a rack in CPU deployment
T_{Pout}	Maximum duration of a power outage that can be managed

to the GCC. This information assists in proper workload management and resource allocation.

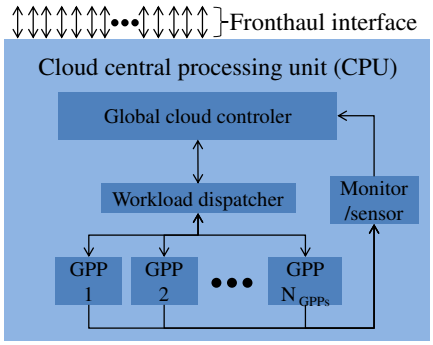


FIGURE 6. Illustration of the edge cloud CPU workload consolidation model [25].

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

$$N^{\text{act}}_{\text{GPPs},t} = \left\lceil \frac{\text{GOPS}^{\text{CPU}}_{t,R}}{\text{CAP}_{\text{GPP}}} \right\rceil, \tag{20}$$

where CAP_{GPP} represents the capacity of the GPP in GOPS, and $\text{GOPS}^{\text{CPU}}_{t,R}$ denotes the required 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 $N_{\text{GPPs}} = \sup_t N^{\text{act}}_{\text{GPPs},t}$.

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

$$s_{\text{CPU}} = \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil s_{\text{rack}}, \tag{21}$$

where CAP_{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

$$P_{\text{CPU},t} = P^{\text{IT}}_{\text{CPU},t} + P^{\text{cool}}_{\text{CPU},t}, \tag{22}$$

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 [26], [27].

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

$$P^{\text{IT}}_{\text{CPU},t} = \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil \text{pw}^{\text{rack}}_{\text{Net}} + P_{\text{GPP},t} N^{\text{act}}_{\text{GPPs},t}, \tag{23}$$

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

$$P_{\text{GPP},t} = \text{pw}^{\text{idle}}_{\text{GPP}} + \left(\text{pw}^{\text{peak}}_{\text{GPP}} - \text{pw}^{\text{idle}}_{\text{GPP}}\right) \frac{\text{GOPS}^{\text{CPU}}_{t,R}}{\text{CAP}_{\text{GPP}}\, N^{\text{act}}_{\text{GPPs},t}}, \tag{24}$$

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 [26], [27].

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 [28]. Consequently, the power consumption of the cooling system in data centers can also be complex to calculate [29]. Despite this, if the cooling PUE is known, the power consumption of the cooling system can then at a time t be calculated as

$$P^{\text{cool}}_{\text{CPU},t} = (\text{PUE}_{\text{cool}} - 1) P^{\text{IT}}_{\text{CPU},t}, \tag{25}$$

where PUE_{cool} is the PUE of the cooling system [27].

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

$$\text{pr}^{\text{CPU}}_{\text{Sinf}} = \left\lceil \frac{P^{\text{peak}}_{\text{CPU}}\, T_{\text{Pout}}}{\text{CAP}_{\text{bat}}} \right\rceil \text{pr}_{\text{bat}} + P^{\text{peak}}_{\text{CPU}} (\gamma_{\text{Co|PD}} + \gamma_{\text{inv}}) + \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil \text{pr}_{\text{rk\&nt}}, \tag{26}$$

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 [26], [30].

IV. COST MODELS

This section presents the cost models utilized to determine the TCO of the UC D-mMIMO system in the context of the methodology in Fig. 3. For the reader's convenience, Table 9 lists all the mathematical representations introduced throughout the equations of this section.

The model is divided into CAPEX and OPEX, which are summed to obtain the TCO. In this context, the CAPEX is given by

$$\text{CAPEX} = C^{\text{CPU}}_{a\&i} + C^{\text{TRPs}}_{a\&i} + C^{\text{Xhaul}}_{a\&i}, \tag{27}$$

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

$$\text{OPEX} = T^{\text{hours}}_{\text{ope}} \left(C^{\text{hourly}}_{\text{fSpace}} + C^{\text{hourly}}_{\text{rep}} + \frac{\text{pr}_{\text{kWh}}}{24} \sum_{n=1}^{N^{\text{daily}}_{\text{samples}}} P^{\text{total}}_{n}\, T^{\text{sample}}_{n} \right), \tag{28}$$

where $T^{\text{hours}}_{\text{ope}}$ is the adopted operational time in hours, pr_{kWh} is the price of kWh, $N^{\text{daily}}_{\text{samples}}$ is the considered number of

TABLE 9. List of mathematical notations introduced in Section IV.

Symbol	Description
$C_{a\&i}^{\text{CPU}}$	Acquisition and installation cost for CPU
$C_{a\&i}^{\text{TRPs}}$	Acquisition and installation cost for TRPs
$C_{a\&i}^{\text{Xhaul}}$	Acquisition and installation cost for fronthaul interfaces
$C_{\text{fSpace}}^{\text{hourly}}$	Hourly costs of floor space
$C_{\text{rep}}^{\text{hourly}}$	Hourly costs of repairs
$F_{l,\text{peak}}$	Peak fronthaul bit rates for TRP l
L_t	Number of active TRPs at time t
M_i	Equipment of type i mean time between failures (MTBF)
N_{TRP}	Number of deployed TRPs
$N_{\text{samples}}^{\text{daily}}$	Number of samples of time during the day
N_i	Number of equipment of type i
N_{tech}^i	Number of technicians required for the repair of equipment of type i
P_n^{total}	Network power consumption at each time sample n
$P_{\text{TRP},t}$	TRP expected power at time t
$P_{\text{Xhaul},t}$	Power associated with the backhaul/fronthaul network at time t
$\mathcal{E}$	Set of different equipment types
$\text{pr}_{\text{FEport}}^{F_{l,\text{peak}}}$	Price of an Ethernet fronthaul switch port capable of supporting rates of $F_{l,\text{peak}}$
pr_{Fdrop}	Price to install the final link from the fiber to the building (FTTB) infrastructure to the TRPs
pr_{GPP}	GPP acquisition price
pr_{kWh}	Price of kilowatt hour (kWh)
pr_{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 small form-factor pluggable (SFP) capable of supporting rates of $F_{l,\text{peak}}$
$\text{pw}_{\text{FEport}}^{F_{l,\text{peak}}}$	Power consumption for an Ethernet fronthaul switch port capable of supporting rates of $F_{l,\text{peak}}$
$\text{pw}_{\text{SFP}}^{F_{l,\text{peak}}}$	Power consumption of a grey SFP capable of supporting rates of $F_{l,\text{peak}}$
S_{tech}	Technicians salary per hour
T_n^{sample}	Duration of each time sample n in the day
$T_{\text{GPP}}^{\text{ins}}$	GPP installation time
$T_{\text{ope}}^{\text{hours}}$	Total operational time in hours
$T_{\text{TRP}}^{\text{ins}}$	TRP installation time
T_{rep}^i	Expected repair time of equipment of type i
T_{trv}	Technicians travel time

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

$$C_{a\&i}^{\text{CPU}} = N_{\text{GPPs}}(\text{pr}_{\text{GPP}} + T_{\text{GPP}}^{\text{ins}} S_{\text{tech}}) + \text{pr}_{\text{Sinf}}^{\text{CPU}}, \tag{29}$$

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

$$C_{a\&i}^{\text{TRP}} = N_{\text{TRP}} \left(\text{pr}_{\text{TRP}} + T_{\text{TRP}}^{\text{ins}} S_{\text{tech}} + \text{pr}_{\text{Fdrop}} \right), \tag{30}$$

where N_{TRP} is the number of deployed TRPs, which is equal to the supremum of L_t. Moreover, $T_{\text{TRP}}^{\text{ins}}$ 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 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 [17], [18], [31]. This work assumes that the fronthaul network utilizes a pre-deployed FTTB infrastructure, a reasonable assumption since the FTTB/fiber to the home (FTTH) penetration is already over 60 % in Europe and east Asia, growing more every year [32], [33]. 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

$$C_{a\&i}^{\text{Xhaul}} = \sum_{l=1}^{N_{\text{TRP}}} \left(2\text{pr}_{\text{SFP}}^{F_{l,\text{peak}}} + \text{pr}_{\text{FEport}}^{F_{l,\text{peak}}} \right), \tag{31}$$

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 $F_{l,\text{peak}}$ [34].

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

$$P_t^{\text{total}} = P_{\text{TRP},t}\, L_t + P_{\text{CPU},t} + P_{\text{Xhaul},t}, \tag{32}$$

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

$$P_{\text{Xhaul},t} = \sum_{l=1}^{L_t} \left(2\text{pw}_{\text{SFP}}^{F_{l,\text{peak}}} + \text{pw}_{\text{FEport}}^{F_{l,\text{peak}}} \right), \tag{33}$$

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 $F_{l,\text{peak}}$ [35].

The hourly repair costs are calculated by

$$C_{\text{rep}}^{\text{hourly}} = \sum_{i \in \mathcal{E}} \left(\frac{N_i\, N_{\text{tech}}^i\, (T_{\text{rep}}^i + 2T_{\text{trv}}) S_{\text{tech}} + \text{pr}_{\text{rep}}^i}{M_i} \right), \tag{34}$$

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 [34].

The hourly floor space costs are calculated by

$$C_{\text{fSpace}}^{\text{hourly}} = \left(s_{\text{rack}} \left\lceil \frac{N_{\text{GPPs}}}{\text{CAP}_{\text{rack}}} \right\rceil + s_{\text{TRP}}\, N_{\text{TRP}} \right) \frac{\text{pr}_{\text{floor}}^{\text{year}}}{8760}, \tag{35}$$

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.

V. NUMERICAL RESULTS

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 UC D-mMIMO systems in the literature, like the one in [3]. 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.

A. CASE STUDY

1) GENERAL ASSUMPTIONS

Fig. 7 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 work 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 (Fig. 8) 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 balance between the fidelity of portraying UE presence and the minimization

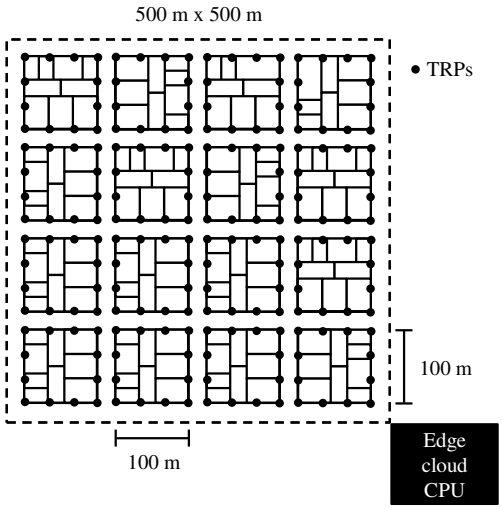


FIGURE 7. Considered urban-dense scenario. TRPs are placed atop buildings at a 15 m height.

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 long-term evolution (LTE) cell at a European metropolitan city [36]. 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 UC D-mMIMO network, which traditionally accounts for 25 % of all UEs [37].

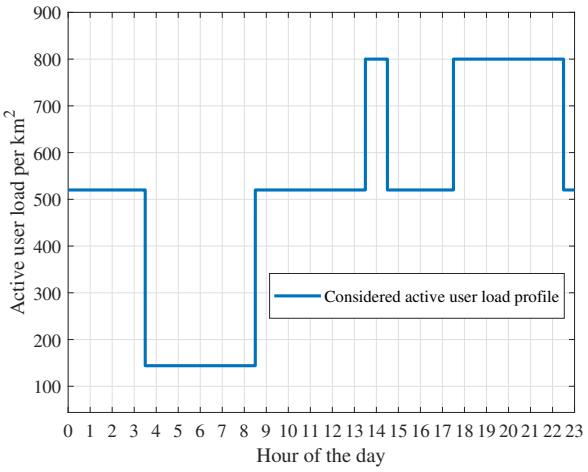


FIGURE 8. Assumed profile of active UEs over the hours of the day.

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 cost unit (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 [38].

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

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 a BCPU functional split. All comparisons focus on the DL 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.

2) SYSTEM MODEL ASSUMPTIONS

The 3rd generation partnership project (3GPP) urban micro (UMi) path-loss model is adopted for the system simulations [39]. 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 Fig. 7. The LoS probability for the calculation of the Rician factor is given by the probability equations in [39] for the UMi scenario. The correlation matrices follow the Gaussian local scattering model [19]. 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 [6].

Table 11 summarizes the system simulation parameters. Most are selected based on parameters commonly adopted in the literature [4], [7], [40], [41]. 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 antennas per TRP (N)	2
Number of supported UEs per TRP ($\max(\lvert\mathcal{D}_l\rvert)$) and UL pilot samples ($\tau_p$)	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 (pw_{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_l^{pil})	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 ($L_{\max}$)	800

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 ($\text{CAP}_{\text{Dcore}}$) is 20 GOPS, the idle DSP core power slope (γ_{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 ($\text{pw}_{\text{DSP}}^{\text{other}}$) is 8.52 W, the price slope for DSP cores (γ_{prDcore}) is 0.42 CU, and the base price of the DSP ($\text{pr}_{\text{DSP}}^{\text{base}}$) 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 [23]. 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.

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

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	-

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 (pr_{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.

4) EDGE CLOUD 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 ($\text{pw}^{\text{idle}}_{\text{GPP}}$) and 652 W at peak ($\text{pw}^{\text{peak}}_{\text{GPP}}$) operation [42]. 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 (CAP_{GPP}) of 7168 for a price (pr_{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 (CAP_{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 ($\text{pr}_{\text{rk\&nt}}$) is 370.4 CU [26].

For the support infrastructure to the IT components, the cooling PUE (PUE_{cool}) is 1.3, while the pricing for cooling and power distribution infrastructure ($\gamma_{\text{Co|PD}}$) is 0.46 CU/W [26], [27]. For the backup power solution, a battery bank is assumed. The acquisition and installation of each battery cost (pr_{bat}) is 11.11 CU, and their capacity (CAP_{bat}) is 1512 Wh [30]. 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 [30].

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 [43]–[45].

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 [46]. Other values are derived from analogous components in different network types [44], [47]. Outdoor fiber MTBF scales with fiber length, which can be obtained as in [48] 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	$\text{pr}^{F_l,\text{peak}}_{\text{SFP}}$
TRP	1	520000	pr_{TRP}
Fr. switch	7	500000	$\text{pr}^{F_l,\text{peak}}_{\text{FEport}}$
Out. fiber	7	$1754386 \times$ km	–
Fiber drop	–	10000000	pr_{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 [44].

B. BASELINE RESULTS

Fig. 9 provides an overview of the TCO after five years of operation concerning the expected UE rate, which is calculated by summing (27) and (28). The analysis includes distributed LP-MMSE and centralized P-RZF and P-MMSE processing implementations under the case study assumptions, as outlined in Subsection V-A. The cost differences between these implementations originate from the variations in the parameters within (27) and (28), which, in turn, are influenced by network requirements calculated in Section III 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.

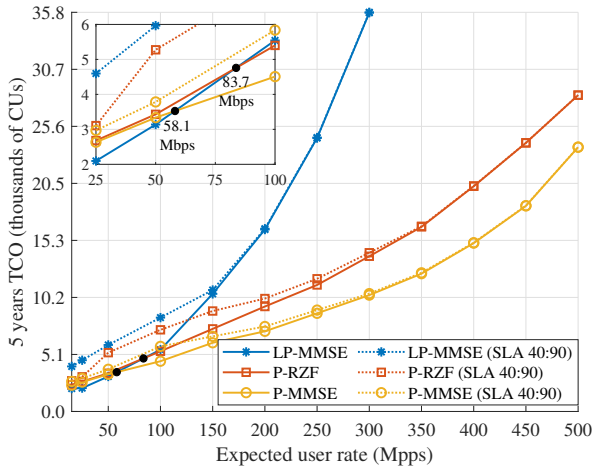
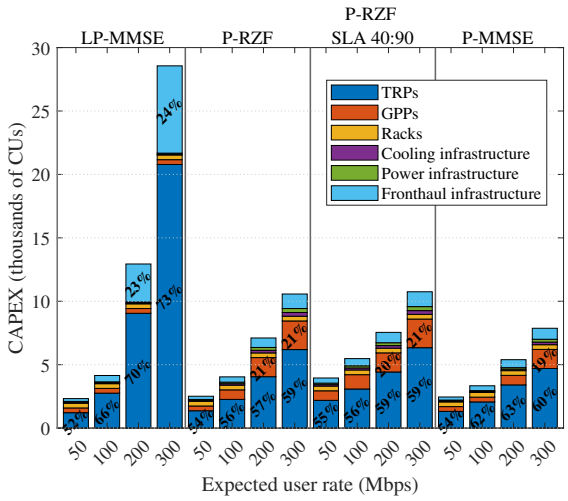


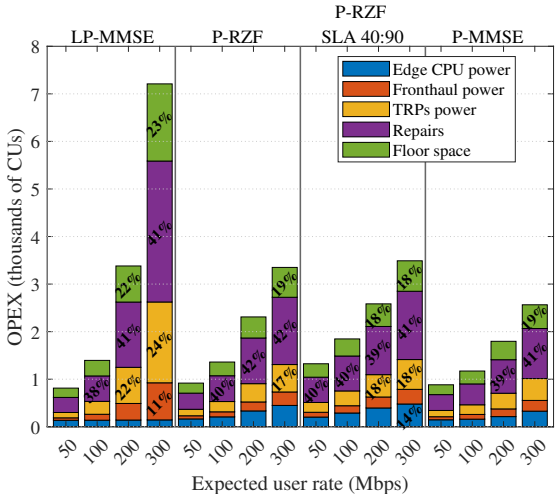
FIGURE 9. TCO after five years of operation concerning the expected UE rate for the case study assumptions ($N = 2$ and $\max(|\mathcal{D}_l|) = 10$). Intersection points between distributed and centralized processing under the same type of TRP deployment are marked by black dots. A zoom of the initial part of the curves is presented at the northwest part of the figure.

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 UC D-mMIMO system, with P-MMSE being the least costly processing alternative.



(a) CAPEX



(b) 5 years OPEX

FIGURE 10. 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.

Fig. 10 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

[2]The required 5G downlink UE rate for an urban wide-area scenario [49]. It can handle Full HD cloud virtual reality (VR) and 4K 3D video [50].

[3]A UE rate capable of handling most bandwidth-intensive applications, such as augmented reality (AR), cloud 2K VR, and 8K 3D video [50].

[4]In 6G systems, improvements should be sought as possible for UL and DL data rates within economic and sustainability constraints, since a 10x or 100x increase from 5G UE rates may be unsustainable [51].

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 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.

Fig. 10a 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 [5].

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.

Fig. 10b 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.

C. 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 UC D-mMIMO network.

1) NON-CPU DEPLOYMENT PRICE REDUCTION

UC D-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 [3]. 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.

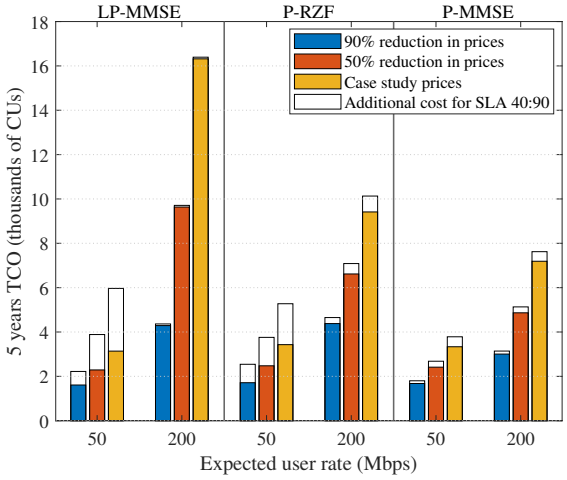


FIGURE 11. 5-years TCO for price variations in TRP and fronthaul prices concerning the case study, including equipment and work-related expenses. The aim is to emulate the potential cost reductions from integrated UC D-mMIMO solutions that reduce installation time and complexity, like the one in [3], 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.

Fig. 11 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 UC D-mMIMO solutions that reduce installation time 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 [3], 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.

2) CPU DEPLOYMENT PRICE REDUCTION

Centralized processing implementations for UC D-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.

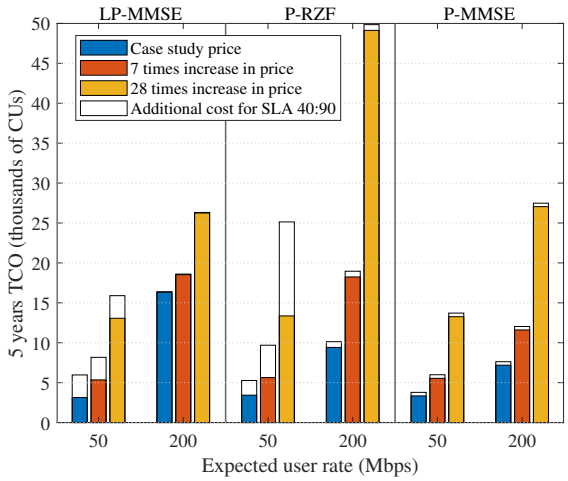


FIGURE 12. 5-years TCO for GPP price variations concerning the case study. Colored bars represent costs for nominal non-SLA TRP deployment, while the colorless stacked bars depict the additional cost incurred by adopting the fairer SLA 40:90 TRP deployment.

Fig. 12 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.

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 work submission. In this context, an analysis of the variation in energy prices is fundamental to ensure that the findings of this work can be applied to different economic realities.

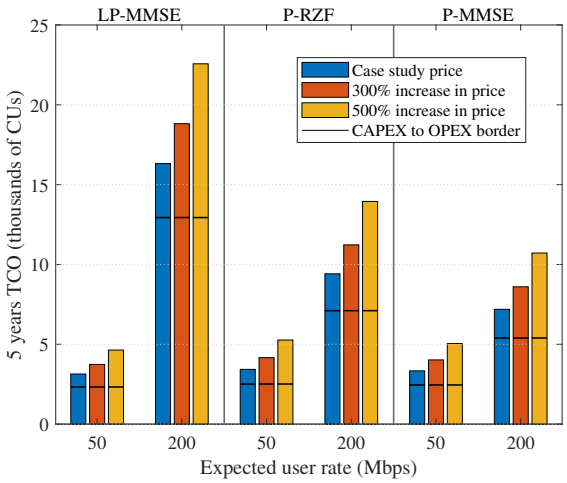


FIGURE 13. 5-years TCO for energy price variations concerning the case study. A line divides the participation of CAPEX and OPEX in the TCO. Only results for the nominal non-SLA TRP deployments are shown.

Fig. 13 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.

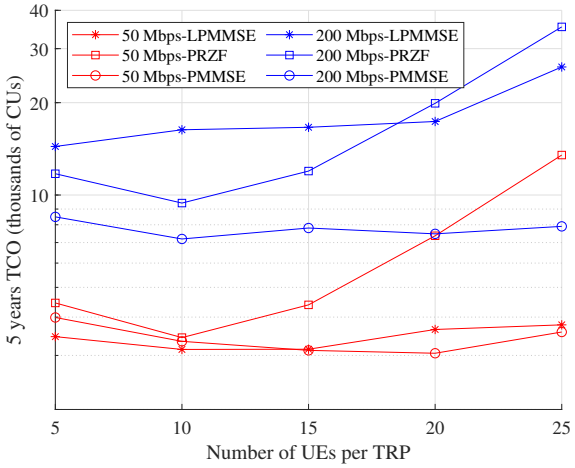
The main change in relation to the case study results is the level of OPEX dominance on the TCO, which becomes much higher as the energy price increases. In fact, OPEX is almost the same as CAPEX for a 500 % price increase over five years of operation. In this situation, extrapolating the results shows 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 UC D-mMIMO systems.

D. IMPACT OF UES SUPPORTED PER TRP VARIATION

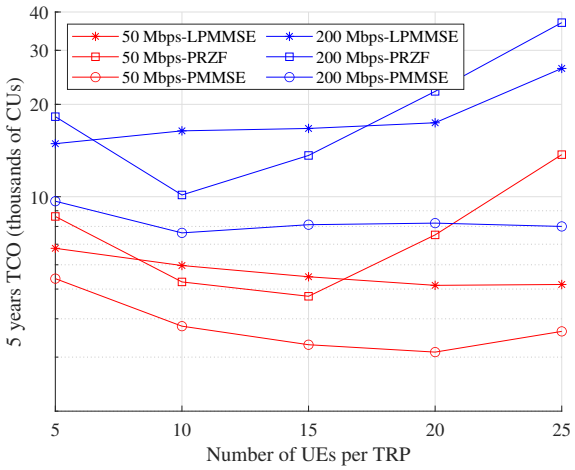
The number of supported UEs per TRP can strongly influence the performance and costs of different processing implementations. For example, the interference levels and computational complexity may experience substantial variations, especially for centralized processing. In this context, analyzing how variations in the number of supported UEs per TRP impact the TCO is essential to make informed decisions regarding processing implementations.

Fig. 14 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.

Fig. 14a 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



(a) Nominal non-SLA TRP deployment



(b) Fairer SLA 40:90 TRP deployment

FIGURE 14. 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.

the optimal 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.

Fig. 14b 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.

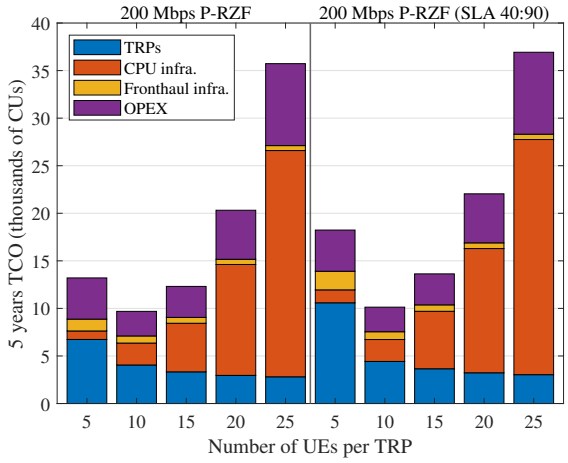
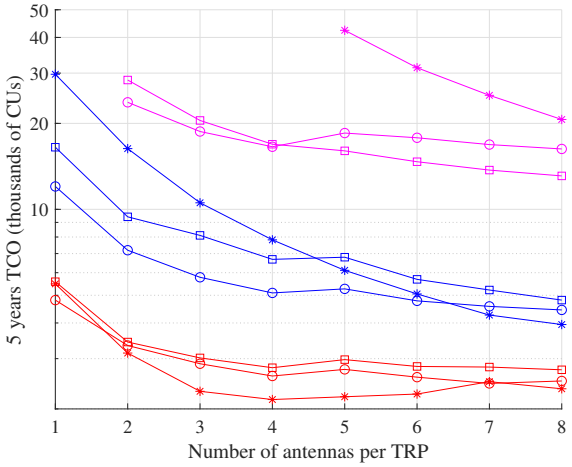


FIGURE 15. TCO composition of P-RZF for demands of 200 Mbps per UE concerning 5, 10, 15, 20, and 25 maximum UEs per TRPs. Other parameters are the same as the case study assumptions. CAPEX is further divided into TRPs, CPU, and fronthaul costs.

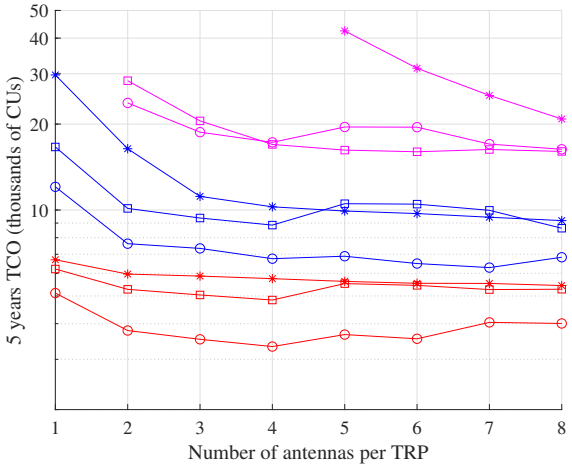
Fig. 15 presents the cost composition of the TCO concerning TRPs, edge CPU, fronthaul infrastructure, and OPEX for the 200 Mbps P-RZF curve in the supported UEs per TRP variation analysis. The aim is to better understand the concave-up behavior of the cost curve. It can be observed that for both Nominal 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.

E. 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.



(a) Nominal non-SLA TRP deployment



(b) Fairer SLA 40:90 TRP deployment

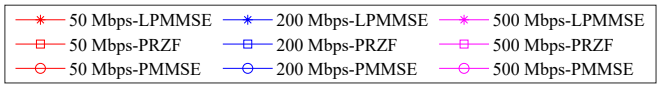


FIGURE 16. 5-years TCO vs. antennas per TRP for 50, 200, and 500 Mbps expected UE rates, representing low, medium, and high demands, respectively. Other parameters remain the same as in the case study.

Fig. 16 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 to this behavior is LP-MMSE under 500 Mbps demands, which already has a TRP count high enough to provide fairness.

Fig. 16a 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.

Fig. 16b 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.

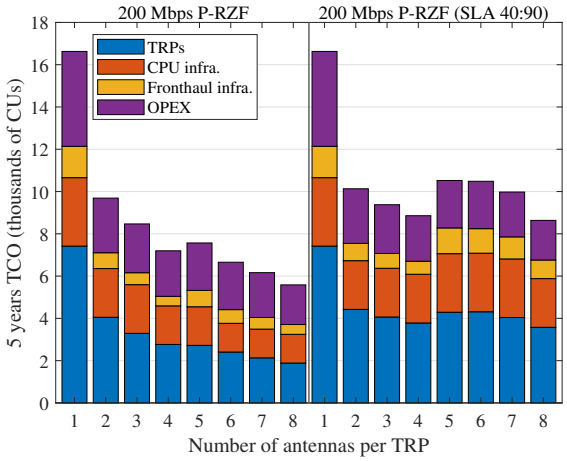


FIGURE 17. TCO composition of P-RZF for demands of 200 Mbps per UE concerning a variation of one to eight antennas per TRP. Other parameters are the same as the case study assumptions. CAPEX is further divided into TRPs, CPU, and fronthaul costs.

Fig. 17 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 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.

VI. CONCLUSION

This paper introduced a comprehensive cost assessment methodology to calculate the TCO of UC D-mMIMO networks. 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. In the latter case, it supports varying UE loads at an expected UE rate, representing the demands from the UE's perspective. This rate is derived from the network-provided average UE rate or a proportional fairness-based UE rate complying with a service level agreement agreed rate. The fairer TRP deployment strategy aims to maintain a significant part of this UE rate throughout a large portion of the coverage area.

The case study carried out in this paper focuses on comparing distributed and centralized processing functional split options on a dense urban deployment. The results, categorized into low, medium, and high demands equivalent to 50, 200, and 500 Mbps per UE, demonstrated that when the TRP deployment does not actively prioritize fairness, distributed processing is more cost-efficient only for low demands. Besides that, a higher TRP antenna count, like eight or more, can make the distributed processing implementation more cost-effective for medium demands. Nevertheless, centralized processing implementation is always more cost-effective for an actively fairer TRP deployment, even for higher antenna per TRP counts.

The analysis of the TCO composition reveals a dominance of CAPEX over OPEX, with TRP costs as the main contributors in centralized and distributed processing implementations. A sensitivity analysis indicates that implementations with reduced fronthaul and TRP deployment costs, with

reduced equipment and work-related costs, have the potential to make distributed solutions more cost-competitive for low and medium demands or at least provide significant cost reductions for all processing alternatives. Further sensitivity analyses suggest that substantially higher-than-normal GPP prices are required to make centralized implementations less competitive. Moreover, a high energy price does not change the cost competitiveness level of the processing alternatives but can strongly reduce CAPEX dominance in TCO.

The evaluation of centralized implementation considered two signal processing solutions: P-MMSE and P-RZF. The results showed that the first was the more cost-effective in low and medium demands. On the other hand, the P-RZF can be more cost-effective in high-demand scenarios 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.

Considering all findings, the centralized implementation utilizing P-MMSE precoding stands out as the most economically viable solution for UC D-mMIMO networks. This approach offers a reasonable cost across various user rates, with the added benefit of becoming even more cost-effective as user data rates increase, thus making it more future-proof than the other alternatives. Furthermore, its costs are less sensitive to the number of UEs served per TRP, avoiding the exponential cost increases in computational complexity and Edge CPU expenses seen with P-RZF. P-MMSE also maintains its cost-effectiveness even with simpler TRPs that have fewer antennas, thanks to its superior interference cancellation capabilities, which reduce the number of TRPs needed compared to distributed scenarios. While certain conditions may make distributed processing or P-RZF precoding more economically feasible, centralized P-MMSE generally offers superior economic benefits.

Finally, building upon the comprehensive analysis conducted in this paper, there are several possible directions for future research. Firstly, it could be beneficial to conduct an analysis with multiple edge CPU, one that includes the signaling dynamics of backhaul links connecting the edge CPU. Secondly, it would be valuable to compare the costs of a UC D-mMIMO setup with a cellular massive MIMO one using distributed and centralized signal processing solutions like single-cell and multi-cell minimum mean square error combining. Thirdly, there is room for further investigation into fronthaul considerations, such as individualizing bit width for data samples in different TRPs, analyzing full greenfield implementation, and considering different access medium and topology options. Lastly, it would be interesting to explore a paradigm shift where the central processing unit is seen as a cloud service hosted within a third-party data center, implying in an edge CPU that the operator does not own.

ACKNOWLEDGMENT

This study was granted access to the computational resources of the High Performance Computer Center (https://www.ccad.ufpa.br/) at the Federal University of Pará.

REFERENCES

[1] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Jun. 2020.

[2] I. F. Akyildiz, A. Kak, and S. Nie, "6G and beyond: The future of wireless communications systems," *IEEE Access*, vol. 8, pp. 133 995–134 030, Jul. 2020.

[3] G. Interdonato, E. Björnson, and H. Quoc Ngo, "Ubiquitous cell-free massive MIMO communications," *EURASIP J. Wirel. Commun. Netw.*, vol. 2019, no. 197, Aug. 2019.

[4] Ö. Demir, E. Björnson, and L. Sanguinetti, *Foundations of User-Centric Cell-Free Massive MIMO*, ser. Foundations and Trends in Signal Processing. Hanover, MA, USA: Now Publ., 2021.

[5] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 1, pp. 77–90, Sep 2020.

[6] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Apr. 2020.

[7] M. M. M. Freitas, D. D. Souza, A. L. P. Fernandes, D. B. da Costa, A. M. Cavalcante, L. Valcarenghi, and J. C. W. A. Costa, "Scalable user-centric distributed massive MIMO systems with limited processing capacity," in *Proc. IEEE ICC*, 2023, pp. 1–7.

[8] G. Femenias and F. Riera-Palou, "Fronthaul-constrained cell-free massive MIMO with low resolution ADCs," *IEEE Access*, vol. 8, pp. 116 195–116 215, Jun. 2020.

[9] L. Furtado, A. Fernandes, A. Ohashi, F. Farias, A. Cavalcante, and J. Costa, "Cell-free massive MIMO deployments: Fronthaul topology options and techno-economic aspects," in *Proc. EuCAP*, 2022, pp. 1–5.

[10] Y. Xiao, P. Mähönen, and L. Simić, "System cost analysis of scalable cell-free massive MIMO architectures for 6G networks," in *Proc. IEEE GC Wkshps*, 2022, pp. 310–316.

[11] Y. Xiao, P. Mähönen, and L. Simić, "Energy and economic efficiency of scalable cell-free massive MIMO networks," in *Proc. PIMRC*, 2023, pp. 1–6.

[12] A. A. Polegre, F. Riera-Palou, G. Femenias, and A. G. Armada, "Channel hardening in cell-free and user-centric massive MIMO networks with spatially correlated Ricean fading," *IEEE Access*, vol. 8, pp. 139 827–139 845, Jul. 2020.

[13] O. T. Demir, M. Masoudi, E. Björnson, and C. Cavdar, "Cell-free massive mimo in o-ran: Energy-aware joint orchestration of cloud, fronthaul, and radio resources," *IEEE J. Sel. Areas Commun.*, Jan. 2024.

[14] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE ICC*, 2019, pp. 1–6.

[15] T. Kim, H. Kim, S. Choi, and D. Hong, "How will cell-free systems be deployed?" *IEEE Commun. Mag.*, vol. 60, no. 4, pp. 46–51, Apr. 2022.

[16] E. J. Oughton and W. Lehr, "Surveying 5G techno-economic research to inform the evaluation of 6G wireless technologies," *IEEE Access*, vol. 10, pp. 25 237–25 257, Feb. 2022.

[17] F. Yaghoubi, M. Mahloo, L. Wosinska, P. Monti, F. d. S. Farias, J. C. W. A. Costa, and J. Chen, "A techno-economic framework for 5G transport networks," *IEEE Wirel. Commun.*, vol. 25, no. 5, pp. 56–63, Oct. 2018.

[18] F. Farias, M. Fiorani, S. Tombaz, M. M. Mahloo, L. Wosinska, J. Costa, and P. Monti, "Cost- and energy-efficient backhaul options for heterogeneous mobile network deployments," *Photonic Netw. Commun.*, vol. 32, no. 3, pp. 422–437, Dec. 2016.

[19] O. Özdogan, E. Björnson, and E. G. Larsson, "Massive MIMO with spatially correlated Rician fading channels," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3234–3250, Jan. 2019.

[20] M. Fiorani, S. Tombaz, P. Monti, M. Casoni, and L. Wosinska, "Green backhauling for rural areas," in *Proc. ONDM*, 2014, pp. 114–119.

[21] H. N. Qureshi, M. Manalastas, S. M. A. Zaidi, A. Imran, and M. O. Al Kalaa, "Service level agreements for 5G and beyond: Overview, challenges and enablers of 5G-healthcare systems," *IEEE Access*, vol. 9, pp. 1044–1061, Dec. 2021.

[22] A. A. El-Saleh, A. Alhammadi, I. Shayea, W. H. Hassan, M. S. Honnurvali, and Y. I. Daradkeh, "Measurement analysis and performance evaluation of mobile broadband cellular networks in a populated city," *Alex. Eng. J.*, vol. 66, pp. 927–946, Mar. 2023.

[23] B. Debaillie, C. Desset, and F. Louagie, "A flexible and future-proof power model for cellular base stations," in *Proc. IEEE VTC Spring*, 2015, pp. 1–7.

[24] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermark, and W. Wajda, "Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," Cordis, EARTH project deliverable D2.3, 2012.

[25] T. Sigwele, A. S. Alam, P. Pillai, and Y. F. Hu, "Energy-efficient cloud radio access networks by cloud based workload consolidation for 5G," *J. Netw. Comput. Appl.*, vol. 78, pp. 1–8, Jan. 2017.

[26] D. Hardy, M. Kleanthous, I. Sideris, A. G. Saidi, E. Ozer, and Y. Sazeides, "An analytical framework for estimating TCO and exploring data center design space," in *Proc. IEEE ISPASS*, 2013, pp. 54–63.

[27] Y. Cui, C. Ingalz, T. Gao, and A. Heydari, "Total cost of ownership model for data center technology evaluation," in *Proc. IEEE ITherm*, 2017, pp. 936–942.

[28] M. Patterson, D. Costello, and P. Grimm, "Data center TCO: A comparison of high-density and low-density spaces," Intel Corporation, White Paper, Jan. 2007.

[29] Y. Zhang and J. Liu, "Prediction of overall energy consumption of data centers in different locations," *Sensors*, vol. 22, no. 10, May 2022.

[30] A. Jahid, M. S. Hossain, M. K. H. Monju, M. F. Rahman, and M. F. Hossain, "Techno-economic and energy efficiency analysis of optimal power supply solutions for green cellular base stations," *IEEE Access*, vol. 8, pp. 43 776–43 795, Feb. 2020.

[31] P. Monti, S. Tombaz, L. Wosinska, and J. Zander, "Mobile backhaul in heterogeneous network deployments: Technology options and power consumption," in *Proc. ICTON*, Jul. 2012, pp. 1–7.

[32] R. Montagne and S. Fogli, "European ftth/b market panorama 2023," FTTH Council Europe, Tech. Rep., 2023.

[33] M. Philpott, A. Fellenbaum, and D. Frey, "Global fiber development index: 2020," Omdia, Tech. Rep., Oct. 2020.

[34] M. Mahloo, P. Monti, J. Chen, and L. Wosinska, "Cost modeling of backhaul for mobile networks," in *Proc. IEEE ICC*, Jun. 2014, pp. 397–402.

[35] M. Fiorani, S. Tombaz, J. Martensson, B. Skubic, L. Wosinska, and P. Monti, "Modeling energy performance of C-RAN with optical transport in 5G network scenarios," *J. Opt. Commun. Netw.*, vol. 8, no. 11, pp. B21–B34, Oct. 2016.

[36] H. D. Trinh, N. Bui, J. Widmer, L. Giupponi, and P. Dini, "Analysis and modeling of mobile traffic using real traces," in *Proc. IEEE PIMRC*, Oct. 2017, pp. 1–6.

[37] Ericsson, "Ericsson mobility report june 2021," Ericsson, Technical Report, 2021.

[38] A. Udalcovs, M. Levantesi, P. Urban, D. A. A. Mello, R. Gaudino, O. Ozolins, and P. Monti, "Total cost of ownership of digital vs. analog radio-over-fiber architectures for 5G fronthauling," *IEEE Access*, vol. 8, pp. 223 562–223 573, Dec. 2020.

[39] 3GPP, "Study on channel model for frequencies from 0.5 to 100 GHz," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 38.901, Mar. 2020, version 16.1.0.

[40] T. Mustala and O. Klein, "eCPRI overview," Ericsson, Huawei Technologies, NEC Corporation and Nokia, Presentation, 2017.

[41] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Elsevier Science, 2020.

[42] 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," Principled Technologies, Inc., Tech. Rep., Oct. 2021.

[43] D. Acatauassu, M. Licá, A. Ohashi, A. L. P. Fernandes, M. Freitas, J. C. W. A. Costa, E. Medeiros, I. Almeida, and A. M. Cavalcante, "An efficient fronthaul scheme based on coaxial cables for 5G centralized radio access networks," *IEEE Trans. Wirel. Commun.*, vol. 69, no. 2, pp. 1343–1357, Feb. 2021.

[44] J. Chen, L. Wosinska, C. Mas Machuca, and M. Jaeger, "Cost vs. reliability performance study of fiber access network architectures," *IEEE Commun. Mag.*, vol. 48, no. 2, pp. 56–65, Feb. 2010.

[45] Principled Technologies, "Reduce hands-on deployment times to near zero with iDRAC9 automation," Principled Technologies, Tech. Rep., Feb. 2020.

[46] C. Di Martino, Z. Kalbarczyk, R. K. Iyer, F. Baccanico, J. Fullop, and W. Kramer, "Lessons learned from the analysis of system failures at petascale: The case of blue waters," in *Proc. IEEE/IFIP DSN*, Jun. 2014, pp. 610–621.

[47] 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," *IEEE Wireless Commun. Lett.*, vol. 11, no. 8, pp. 1580–1584, Aug. 2022.

[48] A. Fernandez and N. Stol, "CAPEX and OPEX simulation study of cost-efficient protection mechanisms in passive optical networks," *Opt. Switch. Netw.*, vol. 17, no. C, pp. 14–24, Jul 2015.

[49] 3GPP, "Service requirements for the 5G system," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 38.901, Mar. 2023, version 19.2.0.

[50] H. GSA and Huawei, "Indoor 5G scenario oriented white paper," HKT GSA and Huawei, White Paper, Oct. 2019.

[51] N. A. e.V., "6G requirements and design considerations," NGMN Alliance e.V., White Paper, Feb. 2023.