O-RAN next Generation Research Group (nGRG)

Contributed Research Report

Use Case Analysis Related to Green Communication in O-RAN

Report ID: RR-2024-08

Contributors:

Dell Technologies

Reliance Jio

Nokia

Mavenir

Rakuten

NVIDIA

NTUST

Release date: 2024.06

Authors

Company	Name	Email
Dell Technologies	Mohammad Alavirad (Editor)	mohammad.alavirad@dell.com
	Hoda Dehghan	hoda.dehghan@dell.com
Reliance Jio	Vikas Dixit	vikas1.dixit@ril.com
Nokia	Daiju Chiriyamkandath Antony	daiju.ca@nokia.com
	Navin Hathiramani	navin.hathiramani@nokia.com
	Ehsan Ahvar	ehsan.ahvar@nokia.com
Mavenir	Sapna Sangal	sapna.sangal@mavenir.com
	Ritesh Parekh	ritesh.parekh@mavenir.com
Rakuten	Kexuan Sun	kexuan.sun@rakuten.com
NVIDIA	Lopamudra Kundu	lkundu@nvidia.com
NTUST	Yu-Chiao, Jhuang	D10902011@mail.ntust.edu.tw
	Setya Widyawan Prakosa	D10702804@mail.ntust.edu.tw
	Tuck-Wai Choong	M11102113@mail.ntust.edu.tw
	Shu-Sheng Wang	M11202150@mail.ntust.edu.tw
	Po-Chun Hung	M11202119@mail.ntust.edu.tw
	Jenq-Shiou Leu	jsleu@mail.ntust.edu.tw

Reviewers

Company	Name	Email
Sumitomo Electric Industries	Hiroshi Miyata	miyata-hiroshi@sei.co.jp
Ericsson	Havish Koorapaty	havish.koorapaty@ericsson.com

Disclaimer

Copyright

Executive summary

6G or beyond networks should target to minimize energy consumption and improve network energy efficiency (EE) as a foundational design goal. It is important to consider energy consumption as well as network EE while adding any new features, as it will contribute to sustainability and environmental protection by reducing the carbon footprint of communication networks.

Future networks are expected to leverage all existing 5G spectrum and in addition, to harnessing new spectrum in the 7-15 GHz range where extreme massive MIMO solutions could be deployed providing high capacity in an energy-efficient manner. Telco cloud is one of the key foundational components of future networks. Designing energy-efficient telco cloud data centers can positively impact overall energy consumption of networks in the 6G timeframe and beyond. Energy optimization can be done in both hardware (e.g., designing energy-efficient servers) and/or in software and algorithms (e.g., resource allocation algorithms).

An objective of this research report is to identify network EE requirements for different network functions, especially in the O-RAN architecture. Network EE requirements both in computation and in communication need to be evaluated. The report discusses advanced design methodologies for EE, including the adaptation of energy-aware protocols and system architectures. It introduces strategies to enhance network EE KPIs, focusing on innovative energy-saving mechanisms and efficient use of resources. It also emphasizes on the necessity of establishing energy-aware design fundamentals and integrating energy-efficient technologies across the network. Furthermore, it elaborates on the critical role of energy-efficient cloud data centers for 6G and subsequent generations, highlighting their influence on the network's overall energy consumption and the importance of employing current and new potential energy-efficient strategies. Additionally, the construction strategies for an energy-efficient AI/ML platform emphasize the integration of sustainable design methodologies and the optimization of energy usage from infrastructure to application services. This research report further suggests ways to reduce energy consumption and promote green communication in networks for 6G and beyond.

Table of Contents

List of abbreviations

3GPP	3rd Generation Partnership Project
5G	5th Generation
6G	6th Generation
ARIMA	Auto-Regressive Integrated Moving Average
ASM	Advanced Sleep Modes
AI	Artificial Intelligence
AR	Augmented Reality
B5G	Beyond 5G
BICSI	Building Industry Consulting Service International
BS	Base Station
CNF	Cloud-Native Network Function
DL	Deep Learning
DT	Digital Twin
E2E	End-to-End
EC	Energy Consumption
EE	Energy Efficiency
ES	Energy Saving
eMBB	enhanced Mobile Broadband
ES	Energy Saving
GPRS	General Packet Radio Service
GPU	Graphic Processing Unit
HRLLC	Hyper-Reliable Low Latency Communications
IMT	International Mobile Telecommunications
IoT	Internet of Things
ITU	International Telecommunication Union
ITU-R	Radiocommunication Sector of the ITU
KPI	Key Performance Indicator
LLM	Large Language Model
LSTM	Long Short-Term Memory
MEC	Multi-access Edge Compute
MIoT	Massive IoT
mMIMO	massive Multiple-Input Multiple-Output
mMTC	massive Machine-Type Communications
ML	Machine Learning

Near-RT	Near Real-Time
NF	Network Function
NFV	Network Function Virtualization
O-RAN	Open Radio Access Network
O-RU	O-RAN Radio Unit
O-DU	O-RAN Distributed Unit
PDCP	Packet Data Convergence Protocol
QoS	Quality of Service
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RSS	Receive Side Scaling
SDN	Software-Defined Networking
THz	Terahertz
TPU	Tensor Processing Unit
UE	User Equipment
URLLC	Ultra-Reliable Low Latency Communications
VM	Virtual Machine
VR	Virtual Reality
VNF	Virtual Network Function
vRAN	Virtual RAN

List of figures

1 Introduction

The growing energy consumption of communication networks is a critical concern as we move towards 6G and beyond. According to reports, information, and communication technologies (ICT) contribute significantly to global greenhouse gas emissions, a figure projected to rise [1]. While emerging technologies like artificial intelligence (AI), machine learning (ML), and virtualization offer opportunities for efficiency, they also introduce additional computational and communication overhead, which can lead to increased energy demands.

An objective of this research report is to identify network EE requirements for different network functions, especially in the O-RAN architecture. Network EE requirements both in computation and in communication of information need to be evaluated. This research report will further suggest ways to reduce energy consumption and promote Green Communication in future networks. In this context, Green Communication refers to network technologies and practices that minimize both energy consumption and overall environmental impact, including greenhouse gas emissions.

Telco cloud will be one of the key foundational components of 6G and beyond. Designing energy-efficient telco cloud datacenters can positively impact overall energy consumption of networks in the 6G timeframe and beyond. It can be done in both hardware (e.g., designing energy-efficient servers) and/or in software and algorithms (e.g., resource allocation algorithms).

This research report first discusses network EE as one of the Key Performance Indicators while designing a system/feature of future networks in chapter 2. Chapter 3 discusses an evaluation methodology to evaluate the performance of the network with respect to EE requirements during the transition from 5G to 6G and beyond. It presents methodologies, frameworks, and case studies for assessing network EE. Additionally, it illustrates the critical role of AI-driven network management in achieving sustainable and energy-efficient future networks, emphasizing the integration of AI and ML to optimize energy consumption. Chapter 4 discusses various strategies to improve the network EE KPIs. Chapter 5 discusses the role of AI models to improve network EE through AI based cross layer strategies. Chapter 6 discusses how information can be communicated from source to target in an energy-efficient manner and the role of AI models in computation and communication. It delves into adaptive resource allocation and cross-layer optimization as key approaches to achieving sustainable and efficient network operations. Chapter 7 discusses construction strategies for developing an energy-efficient AI/ML platform. Chapter 8 discusses the very important topic of network EE in telco cloud. And finally, chapter 9 presents the conclusions of this research report.

2 Network EE as KPI for future network system/features

Network EE has been identified as a foundational design goal for next generation networks, and the expectation is that its importance will continue to increase in future generations. To achieve improved network EE, it is essential to develop advanced network EE metrics for diverse next generation use cases such as the Internet of Things (IoT), massive Machine-Type Communications (mMTC), and Ultra-Reliable Low Latency Communications (URLLC), providing a comprehensive evaluation of network performance. Next generation systems hence should just be the steppingstone towards a native energy-efficient system with future generations improving over the initial baseline in terms of network EE. To ensure this increasingly important design goal is successfully accomplished, one could envision the need for EE requirements to guide the design and EE metrics to evaluate the progress. Further, refining algorithms for energy optimization, focusing on dynamically adjusting network parameters based on real-time energy consumption data, is crucial. Different evaluation metrics would be required for different stages of the development (specifications vs implementation vs deployment). These metrics would provide the ability to deploy closed loop feedback, where learnings can be employed for future designs and deployments. Additionally, expanding KPIs to include metrics reflecting the societal and environmental impact, like the network's carbon footprint and the proportion of energy sourced from renewables, is essential.

The importance of network EE is further reflected in the IMT-2030 future technology report (ITU-R M.2516-0 [2]). This report states that IMT-2030 capable systems will provide enhanced data rate, mobility, spectrum efficiency, latency, reliability, connection density, network EE, and area traffic capacity to efficiently support emerging usage scenarios and applications. This should include a lifecycle analysis of energy consumption across different network elements, emphasizing the total environmental impact. The IMT-2030 report further stresses that key drivers to achieve at least environmental sustainability should include network EE, sensing resolution and accuracy, a simplified user centric network, and native Artificial Intelligence (AI). From the network EE perspective, it states how the use of power efficient technologies, both in backhaul and access, can facilitate the adoption of small-scale renewable energy sources.

The expectation is that IMT-2030 will proceed to include RAN performance requirements to guide an energy-efficient design, however, once systems are designed and deployed there is a need to be able to monitor the efficiency gains achieved. IMT-2030 minimum performance requirements are RAN centric and do not provide requirements for subsystems outside the radio access network.

To ensure energy-centric design principles will also be applied in O-RAN for 6G and beyond, O-RAN specific network EE principles and requirements could be established. These principles and requirements need to complement 3GPP technologies and be specified on the basis of the fundamental O-RAN architecture to ensure, for example, sustainable next generation O-RAN automation, management, orchestration, and optimization. Although the O-RAN radio unit (O-RU) is the primary consumer of energy within an O-RAN network, design principles should not focus on establishing hard

limits on energy consumption for the O-RU(s). It is important to not restrict design and innovation by not considering, e.g., bands supported by the O-RU, their time spent in each power state or antenna configurations. O-RU energy consumption should be evaluated considering product capabilities. O-RAN energy-centric requirements should instead focus on requirements to, for example, ensure:

- sustainable AI/ML deployments
- exposure of energy information from different network elements or subsystems or xApps/rApps to enable energy-efficient orchestration and management.
- xApps/rApps providing estimated energy consumption profiles to enable Near-real-time (RT) RAN Intelligent Controller (RIC) or Non-RT RIC to select adequate feature activation based on a desired energy consumption.
- architectures to enable shared data pools employed by AI/ML algorithms deployed in xApps or rApps or other network entities.

The O-RAN EE requirements could further be supported with evaluation methodologies to provide some coarse understanding of a feature's impact on network EE. Here, the term coarse is employed to reflect that features typically do not exist in isolation, and hence interworking with other features or architectural aspects is paramount to determine the actual network EE during deployments.

There have been many KPIs specified to measure the network EE of a 3GPP system. The 3GPP KPIs specified to measure network EE of a 5G network and slice are captured in 3GPP TS. 28.554 [3]. From a RAN perspective, the primary network EE KPIs defined are centered around the enhanced Mobile Broadband (eMBB), URLLC, and MIoT use cases and are based on the total uplink and downlink PDCP data volume transacted by the RAN with respect to the energy consumption of the network elements within the RAN for a defined period of time. This bits per Joule metric can become distorted when the RAN is used for other use cases where spectral efficiency is not the prime concern. For example, for IoT use cases, coverage and connection density are fundamental, as is latency for URLCC use cases. Although specific variations of the bits/Joule formulation exist for these use cases, including weighted versions to encompass more than one use case, it is difficult to visualize the network EE of the RAN with a single KPI. The challenges stated with network EE KPIs do not allow to benchmark the network EE of RAN cluster within or between networks but do allow to observe trends and these trends could provide the initial seed for further optimizations.

The network EE KPIs become even more obscure the deeper one observes in the network due to centralization and pooling aspects. These aspects increase the complexity to discern the amount of energy consumed by a specific function for a specific RAN node. Capabilities to provide granular energy consumption measurements based on methods such as Kubernetes Efficient Power Level Exporter (KEPLER [4]) could improve the accuracy and interpretability of these end to end (E2E) KPIs.

Future work should not only focus on enhanced multi use case network EE KPIs for benchmarking but also englobe KPIs for carbon emissions and key value indicators for societal impact. Metrics such as those considering the energy source and carbon emissions of the energy employed by a network element (e.g., amount of renewable energy) could be further employed during orchestration decisions. Societal impact should be captured as a key value indicator based on observations of not only the energy and carbon footprint of a network but also its enabling effects.

If O-RAN specific network EE requirements are established, a set of corresponding O-RAN specific network EE KPIs should be considered. As stated above, due to the nature of centralization, pooling and adoption of multipurpose nodes, it may not be possible to define a perfect KPI. However, this too can be fine-tuned based on learnings and correlations observed from field deployments.

3 Methodology to evaluate the network EE requirement

Amidst the rapid evolution of technology and the escalating demands for high-performance communication networks, the emergence of the next generation of mobile networks introduces transformative capabilities. However, this progression necessitates a focused examination of a critical facet, network EE. This chapter endeavors to establish a robust methodology for the systematic evaluation and enhancement of network EE for the future within the domain of networks.

The envisioned next generation network is not only a technological entity but a complex ecosystem where sustainability must be a foundational pillar. As such, this chapter seeks to clarify the multifaceted nature of network EE in 6G. We will delve into the details of energy consumption across diverse next generation applications, from high-bandwidth, low-latency services to massive IoT deployments, and the essential role of AI in orchestrating an energy-efficient and sustainable network architecture. This involves exploring how AI models, including those developed using emerging training strategies like federated learning and transfer learning, can contribute to optimizing energy usage in various communication scenarios [10]. By understanding the current energy consumption trends, exploring the potential of AI and ML in managing energy demands, and examining various future use cases, we aim to contribute to the development of a sustainable, green communication paradigm in the 6G and beyond era [5], [6].

3.1 5G vs 6G: network EE using new intelligence techniques

As we transition from 5G to 6G, the landscape of network infrastructure and connected devices is expected to undergo an exponential increase. This surge, while promising in terms of connectivity and service quality, poses a significant challenge in terms of energy consumption. The information industry, now more than ever, is under pressure to mitigate energy overheads and reduce reliance on fossil fuels. This is particularly crucial given that 5G base stations (BSs) and mobile devices are already consuming markedly more energy than their 4G predecessors. For instance, a typical 5G BS with multiple frequency bands can consume over 11,000 watts, a stark contrast to the less than 7,000 watts consumed by a 4G BS [7].

The increased power consumption in 5G, and by extension in next generation, mainly stems from two factors: the growing power for amplification in massive Multiple Input Multiple Output (MIMO) antennas and the processing of burgeoning data [8]. This shift indicates that while we have achieved a reduction in energy consumption per unit of data, the overall energy demand remains on an upward trajectory [9]. Therefore, in the future, where we foresee an even more intricate and expansive network infrastructure, addressing these energy demands becomes imperative.

Furthermore, the integration of AI in managing networks and improving network EE presents a novel approach towards green communications. AI techniques, especially advanced ML methods like Deep Learning (DL), are extensively recognized as potent tools to mitigate energy demand, enhance network EE, and manage the dynamic

energy harvesting process expected to be widely adopted in 6G [10], [11]. Additionally, there is a specific energy consumption component related to in-network AI/ML functionalities/LCM, which is also part of the energy "cost" that should be reduced.

3.2 AI-enabled network EE framework in 6G and beyond

The conceptualization of network EE in the context of future networks revolves fundamentally around optimizing the domain KPI (e.g., data rate in communication domain measured in bits/sec) with consideration given to power consumption (measured in Joules/sec), often articulated as 'bits per Joule'. However, it's important to note that energy consumption in the communication domain extends beyond data transmission and reception events. It also encompasses energy consumed during data storage and processing, such as transformation through compression algorithms. This metric becomes increasingly critical as we advance into the future era, where network densification, higher frequency bands, and more complex (and advanced) modulation schemes are expected to significantly augment energy demands. Network EE in any domain can be mathematically expressed as $\eta_{domain} = \mathrm{KPI}_{\mathrm{domain}}/\mathrm{Consumed\ power}$, for instance, RAN EE is defined as:

$$\eta_{Comm.} = \frac{\mathrm{Data\ Rate}_{(\mathrm{bits/sec})}}{\mathrm{Power\ Consumed}_{(\mathrm{Joules/sec})}}$$

This formula serves as a fundamental KPI in evaluating and designing networks for 6G and beyond. It covers the dual objectives of maximizing data throughput while minimizing energy expenditure. Let’s discuss on integration of AI/ML in next generation for network EE: AI/ML techniques, particularly DL, have been identified as critical in transforming 6G and beyond networks into more energy-efficient systems [12]. These technologies offer detailed insights and predictive capabilities which are essential for dynamic network optimization [12]. The application of AI/ML for network EE can be categorized into three main areas:

- **DL applications for network parameter optimization**: DL excels in fine-tuning network parameters to enhance energy efficiency. It can be used for traffic pattern analysis, predicting network loads, and pre-emptively adjusting network parameters to optimize energy usage [13]. Implementing DL algorithms for intelligent resource allocation can facilitate optimal usage of network resources with minimal energy wastage [13].

- **AI-driven network management**: While DL focuses on optimizing network parameters, AI-driven management takes a broader approach, addressing the physical infrastructure of the network. AI can dynamically manage BS operations, including switching BSs on/off based on user density and traffic demands, significantly reducing energy consumption. AI algorithms can adjust UE parameters for energy conservation without compromising on service quality [12].

- **Emerging training strategies:** Federated Learning (FL) involves distributed training of AI models. This method reduces energy consumption at central servers and enhances the efficient utilization of edge resources. Furthermore, studies have shown that FL can lead to a reduction in overall or total energy consumption when compared to traditional centralized training approaches [14]. Another strategy, Transfer Learning, accelerates AI model training by applying knowledge gained from one problem domain to a different but related domain, thereby reducing computational energy requirements.

These categories work synergistically, with DL providing the analytical foundation for parameter optimization, AI-driven management making higher-level decisions about network infrastructure, and emerging training strategies enhancing the efficiency of the AI/ML processes themselves. Together, they form a comprehensive approach to maximizing network energy efficiency in 6G and beyond, addressing both immediate operational efficiencies and long-term strategic planning.

Let’s look at an example of a network workload management use case in networks for 6G and, where the penetration loss in some frequency bands with higher frequency radio signals limits BS coverage, contributing to increased energy consumption. Uneven user distribution and mobility create imbalanced traffic loads. From network EE perspective, minimizing active BSs is proposed through careful deployment, workload management, and user association. Dynamic scheduling of multi-tier BSs based on changing network traffic from user mobility aims to reduce energy consumption while maintaining a qualified connection and meeting Quality of Service (QoS) requirements [10]. Leveraging user movement patterns and historical traffic data, ML models like Auto-Regressive Integrated Moving Average (ARIMA), random forest, Long Short-Term Memory (LSTM), and ensemble learning can be employed to predict traffic changes. Decisions to switch BSs on or off are based on predefined KPI thresholds, as detailed in other research which uses neural networks for traffic prediction [10]. A balance between coverage and efficiency loss is achieved through adjustable thresholds, with simulations showing a potential energy consumption reduction of up to 63% while satisfying over 99.9% of service requests. Figure 3-1 illustrates how BS switching strategies can be implemented.

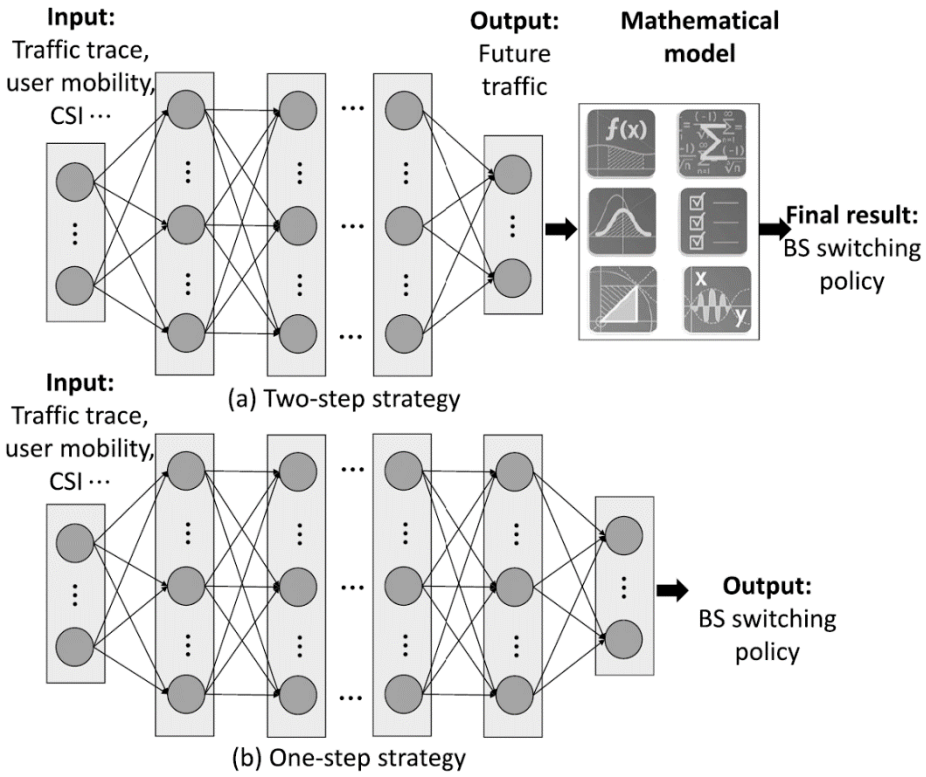

Figure 3-1 One-step and two-step AI-based BS switching strategy [10]

In the next sections, we delve into use case analysis of network EE in the future, but it is notable to know that the integration of AI/ML into networks for 6G and beyond for optimizing network EE is not just a technical enhancement but a paradigm shift. By leveraging these technologies, future networks can go beyond traditional network management limitations, paving the way for a more sustainable, intelligent, and energy-efficient future. This theoretical framework sets the foundation for a detailed exploration of specific use cases, methodologies, and practical applications in the subsequent sub-chapters, all aimed at actualizing the vision of an energy-efficient 6G network.

3.3 B5G/6G verticals and network EE implications

In the pursuit of B5G/6G verticals and their network EE implications, we explore challenges and innovative solutions within the evolving landscape of 6G and beyond networks in some important verticals:

3.3.1 Immersive communications

In 6G and beyond, the immersive communications use case prioritizes delivering significantly higher data rates and capacity compared to 5G. This emphasis on elevated performance is particularly crucial for applications demanding substantial bandwidth, including virtual reality (VR), augmented reality (AR), and ultra-high-definition video streaming [15].

However, while pursuing high data rates in immersive communications, it is crucial to address challenges related to overall network energy consumption to mitigate the associated increased power demands on network infrastructure—especially in BSs and data centers. Although the network energy consumption per bit is expected to improve in 6G compared to previous generations, the massive increase in data traffic and network complexity may lead to higher total energy consumption if not carefully managed. The deployment of advanced MIMO techniques and beamforming in immersive communications further intensifies these energy demands [16]. To address these challenges, potential solutions include the implementation of AI-driven resource allocation and the adoption of network slicing to optimize energy usage within immersive communications [17]. Additionally, efforts towards developing energy-efficient hardware components, such as low-power amplifiers and antennas, are deemed essential for mitigating the energy impact of immersive communications applications on 6G and beyond networks [18].

3.3.2 Massive communications

Massive communications in next generations involves the connection of a diverse range of IoT devices, including sensors, actuators, and smart home appliances, necessitating low power and bandwidth while requiring ubiquitous connectivity [19]. The energy implications of massive IoT are substantial, as the sheer volume of connected devices can result in significant energy consumption for both the devices themselves and the supporting network infrastructure [20]. To address this, strategies for network EE are crucial. This includes the adoption of energy harvesting technologies, allowing devices to derive energy from their environment (such as solar, thermal, or kinetic energy). Additionally, implementing low-power communication protocols and utilizing edge computing to reduce data transmission distances emerge as effective measures to greatly enhance network EE [21].

3.3.3 Hyper-reliable low-latency communication (HRLLC)

In the context of HRLLC, which is designed for applications demanding high reliability and low latency, such as autonomous vehicles, industrial automation, and telemedicine, network EE is an essential consideration [22]. The challenge lies in the fact that achieving low latency and high reliability often necessitates dense network deployment, rapid data processing, and transmission redundancy (e.g., in the form of retransmissions), leading to increased energy usage [23]. To address these concerns, mitigation approaches are explored. Optimizing the network performance through the application of AI/ML techniques emerges as a promising strategy to reduce energy consumption in HRLLC scenarios [24]. Implementing techniques like predictive analytics for traffic management and intelligent sleep modes for network equipment can contribute to conserving energy and making HRLLC systems more sustainable [25].

3.3.4 Energy harvesting techniques for IoT and EE considerations for high frequency bands

Energy harvesting will play an important role in 6G and beyond, particularly for IoT devices. Various techniques, including ambient RF energy harvesting, photovoltaic cells, and piezoelectric energy generation, can offer sustainable energy sources for

low-power devices [26], [27]. In addition, future network may utilize high frequency bands, up to Terahertz (THz) frequency bands (0.1-10 THz) for ultra-high-speed data transmission [28]. Despite providing tremendous data rate capabilities, high frequency communications present challenges such as high-power consumption and significant signal attenuation [29]. The integration of THz bands in 6G and beyond networks will require the development of energy-efficient THz transceivers and amplifiers. Furthermore, optimization of beamforming and MIMO technologies in THz communications is essential to strike a balance between high-speed data transmission and network EE considerations [30].

Building upon the foundational IMT-2020 service types—eMBB, mMTC, and URLLC—we may delve into their evolution into next generation scenarios as outlined in ITU-R M.2160 [31]. These include immersive communication, an advanced form of eMBB that delivers rich, interactive video experiences; massive communication, which expands upon mMTC by connecting an even greater number of devices or sensors for a wide range of applications; and HRLLC, an enhanced version of URLLC with more stringent requirements on reliability and latency for specialized use cases. Additionally, new usage scenarios have been introduced: Ubiquitous Connectivity aimed at bridging the digital divide, AI and Communication supporting distributed AI applications, and Integrated Sensing and Communication combining sensing capabilities with communication. These scenarios are designed to address emerging trends and overarching principles such as sustainability, security, and ubiquitous intelligence for 2030 and beyond [32]. These expanded and new use cases pose unique challenges to network EE, necessitating the development of innovative solutions and methodologies for systematic evaluation and enhancement. Our focus should remain on striking a balance between harnessing the transformative capabilities of next generation networks and ensuring network EE, a critical facet in the rapidly evolving landscape of communication networks.

Beyond the key use cases discussed, other next-generation network applications present unique energy and efficiency challenges. Addressing these challenges requires integrating AI/ML, adopting energy harvesting technologies, and developing energy-efficient components. As networks evolve, continuous innovation will be key to balance network EE with performance and connectivity goals.

3.4 Methodology for enabling network EE

The methodology for enabling network EE in networks for 6G and beyond involves two key components: data collection and analysis techniques.

3.4.1 Methods of data collection:

Data collection in networks for 6G and beyond for network EE evaluation can be achieved through various methods, such as network monitoring tools and probes (that can capture real-time network performance data, including traffic loads, BS operational statuses, and network usage statistics), user device data collection (information about mobility patterns, application usage, and network conditions experiences by end-users), and network logs and records (maintained by network element management

systems which can serve as source of historical data on network usage, traffic patterns, and operational events). The data collected through these methods may include, but is not limited to:

- **Network usage statistics***:* The cornerstone of network EE evaluation lies in gathering detailed network usage statistics. This includes tracking metrics such as data rates which can vary from a few kbps in low-power IoT devices to several Gbps in high-speed connections, bandwidth utilization percentages, and detailed signal quality indicators like Signal-to-Noise Ratios (SNR). Advanced network monitoring systems are deployed, capable of processing terabytes of data, to capture these metrics and evaluate them. The systems are designed to identify patterns in data usage, such as peak hours where data rates may surge up to 50% higher than average [33].

- **User mobility data and pattern analysis**: Understanding user mobility is essential for optimizing network resource allocation. This involves collecting raw data such as current location data from sources like GPS and cell tower connections, as well as historical location data. The collected raw mobility data is then analyzed to derive meaningful patterns. For example, data showing a 30% increase in user density in certain urban areas during rush hours can help in pre-emptive network resource allocation, thus enhancing network EE. ML models, particularly those specializing in pattern recognition like Convolutional Neural Networks (CNNs), are employed to process and analyze the collected mobility data to extract these patterns [33].

- **Traffic loads analysis**: Monitoring traffic loads involves more than just quantifying the data. It requires a breakdown of traffic types – such as a 40% share of video streaming traffic or a 25% share of IoT communication – and their respective energy implications. Traffic analysis tools are employed to categorize data flows, and their findings are crucial in deciding which parts of the network require energy optimization [33], [34].

- **BS operational statuses**: Each BS's operational status, whether it is in active mode consuming, for example, 800 to 1500 watts, idle, or switched off, is continuously monitored. This data is vital, especially considering that a BS in idle mode can reduce its energy consumption by up to 60% compared to its active state. Cell level statuses, such as the operational state of mMIMO bands/carriers, are also essential to monitor, as they indicate whether these crucial components are functioning correctly. Remote monitoring systems integrated with AI analyze this data to provide insights on optimal BS operational strategies, like reducing transmission power by 20% during low-traffic periods without affecting network performance [35].

3.4.2 Analysis techniques:

This includes time-series analysis for traffic patterns and predictive modeling for anticipating future network demands. The core of network EE analysis lies in the effective processing and interpretation of vast amounts of collected data. This includes network usage statistics, user mobility and behavior patterns, traffic load distributions, and operational statuses of network components like BSs and routers. Various ML models are employed to analyze this data. These models range from classical statistical approaches to more advanced DL techniques, each suited for different types of data and analysis requirements.

- **Time-series analysis**, which is essential for understanding and predicting network traffic dynamics. This involves studying data collected over time to identify trends, seasonal variations, and cyclical patterns in network usage [36]. Models like ARIMA, LSTM networks, and other sequence prediction models are used. These models help in forecasting future traffic scenarios based on historical data, allowing network operators to pre-emptively adjust network configurations for optimal energy usage [37]. For instance, a time-series analysis might reveal that network traffic peaks on weekday evenings and drops significantly overnight. This insight can guide the scheduling of BS operational intensities and, consequently, the optimization of energy consumption.

- **Predictive modeling**, which extends beyond short-term traffic forecasts. It encompasses the anticipation of long-term network growth and evolving user behaviors [38]. This is particularly important in the context of next generation network, where new technologies and user applications are continuously emerging. The integration of AI with big data analytics allows for the creation of models that can predict how network demands will evolve with changes in technology, user behavior, and even socio-economic factors [39], thus enabling energy-saving adjustments in the network's operation [40]. For instance, predictive models can be used to simulate the impact of introducing new IoT devices into the network, providing insights into how this will affect energy consumption and efficiency [41].

When analyzing techniques, it is essential to prioritize model accuracy and computational efficiency. Real-time data processing capability is also crucial for effective analysis. This allows for immediate adjustments to network operations, ensuring continuous optimization of EE, and feedback loop for continuous improvement which enables the models to adapt to changing network conditions and maintain high accuracy over time.

3.5 AI-based energy-efficient network management

In the realm of energy-efficient network management, particularly within advanced cellular networks like 6G, several strategies are employed to optimize energy consumption while maintaining high levels of service quality. These include the

deployment of small cells, which can potentially enhance coverage and capacity in densely populated areas and may contribute to overall EE depending on the specific deployment scenario. Network slicing is another approach, allowing for the creation of multiple virtual networks on the same physical infrastructure, each optimized for specific types of traffic and services, thus potentially enhancing overall network EE. Additionally, the use of advanced sleep modes in network equipment helps in reducing power consumption during low traffic periods. Beamforming technology, which directs signals precisely to the intended users, may contribute to network EE by minimizing signal wastage, although it is important to note that it can also increase power consumption due to additional processing requirements. Among these strategies, user association stands out as a critical component, playing an essential role in balancing network load and reducing energy consumption. By intelligently connecting user devices to the network nodes that can provide services with minimal energy consumption while meeting performance requirements, user association not only optimizes the user experience but also contributes to the overall network EE of the network. This aspect of network management, particularly in the context of next generation networks, warrants a more detailed exploration.

Unlike static rules, AI facilitates real-time, dynamic user-BS associations. These associations adjust according to network conditions and anticipated changes, aiding in the balance of network loads and reduction of energy consumption [42]. In terms of techniques and models in AI for user association, ML models such as Reinforcement Learning (RL) and Deep Neural Network (DNN) are employed. These models are proficient at learning and implementing optimal user association strategies, demonstrating a strong ability to recognize and predict patterns in network usage [43].

AI models in next generation network face challenges of massive connectivity and service heterogeneity. To address these challenges, these models are designed to be scalable and operate with minimal latency, ensuring timely and efficient network management [44]. Case studies and simulations have shown that AI models can dynamically adjust user associations in urban areas for optimal energy use. Furthermore, AI-driven policies dynamically adjust user-BS associations based on current network conditions, user mobility, and traffic demands. This ensures optimal energy usage while maintaining service quality.

4 Strategies for improving the network EE KPI

The RAN is responsible for the majority of the Energy Consumption (EC) of a mobile network, and the O-RU consumes the largest part of the energy consumption of the RAN [45]. The rarefication of fossil fuel-based energy resources and the urgent need to reduce CO_2 emissions make energy saving a strategic goal for network operators, in addition to the significant energy related Operating Expense (OPEX) reduction requirement.

Network EE can be improved using various approaches. On cell level, several energy saving (ES) mechanisms are related to switching off certain components in the O-RU to save energy. In a longer time-scale of minutes, hours and above, and when the cell load is low, ES can be achieved by switching off one or more carriers or the cell itself. At the same or shorter time-scale, from seconds to minutes, ES can be achieved by switching off RF channels (including possibly array of antennas) of a Massive-MIMO (mMIMO) system. At a shorter time-scale corresponding to a symbol, subframe and frame, Advanced Sleep Modes (ASM) can be applied to switching off and on O-RU components quickly on per symbol, per subframe or per frame basis.

The energy saving technologies by switching off O-RU components is already made available in 5G with several enhancements to existing 5G specifications. In the next generation, air interface protocol needs to be designed for EE from the very beginning considering both performance optimization, spectrum efficiency and EE optimization. Similar to today how the air interface protocol adapts to very different service requirements (e.g., eMBB, URLLC, eMTC), improvement can be done to be more adaptive to different performance and energy saving requirements in different scenarios. More cross layer optimization with AIML enhanced intelligence can further optimize the EE and find the best balance point between performance and energy saving. Computing resource for processing workload of both networking and application services can be jointly optimized to further improve EE of the end-to-end system.

Apart from achieving energy saving by switching off O-RU hardware component when it is not used, more energy-efficient O-RU and O-DU hardware can be adopted. The EE should become an important KPI to measure for each vendor’s product, which motivate the innovation and adoption of more energy-efficient RF and computing hardware in the O-RU and O-DU from all vendors. More efficient and greener power supply source can also be leveraged, e.g., using solar and wind power for O-RUs and O-DUs at the cell site and edge data center.

On system level, EE can be improved from radio planning aspect. Along with planning the radio network to optimize the coverage, interference, and capacity, minimizing radio power waste is another important factor to be considered. High gain antennas, higher dimension mMIMO, smart reflectors, UE centric network are effective techniques that can be used to maximize coverage, system capacity while minimizing interference and radio power waste. Digital twin (DT) and AI/ML can also be leveraged to automatically derive optimal solutions considering factors and goals in various

orthogonal dimensions. The energy consumption of the DT and AI/ML models themselves also need to be accounted and ensure a net gain can be achieved.

On the network topology planning side, more centralization of baseband processing can also significantly reduce energy consumption. For network slices providing non-delay critical service only, baseband processing can be done in the central data center which benefit from the centralization gain (computing power is allocated on-demand rather than reserved statically for each cell which result in more energy waste). Baseband processing for low latency slices can be done at the edge data center for lower latency but most of the cases only need to provide relatively lower throughput capacity which result in lower energy consumption at the edge data center as well. Aggregating data stream from multiple network slices can be done at the O-RU level rather at the O-DU today for more efficient energy usage.

On the cloud computing side, more efficient cloud computing hardware can be adopted. In addition, more intelligent, dynamic, proactive, and cross layered energy management and optimization mechanisms can be developed. They can dynamically and accurately adjust CPU, GPU, FGPA and ASIC power state, frequency scaling, core pinning and switching off unused hardware components etc. for the most efficient use of hardware in various traffic conditions. More optimized and adaptive mixing of the computing platforms for different types of processing workloads varied in different traffic conditions is another important method to optimize the EE at the cloud data center. More computing layer abstraction can be achieved in next generation to enable flexible and adaptive computing hardware selection for most efficient use of computing hardware.

5 AI/ML models to improve the network EE

Wireless systems are designed to support the peak capacity at any time. But in real deployments, peak conditions are observed only at certain duration. Most of the time, system works at medium capacity or at very low capacity.

There can be multiple methods to reduce the energy consumed by a wireless system depending on the loading conditions.

- Network energy saving optimization by Non-RT RIC/ Near-RT RIC. Response time could be in seconds or in milli seconds.
- Energy consumption reduction mechanism within the NFs. Response time could be controlled to even slot level.

Numbers of cores assigned to successfully run the RAN NFs on the server/product are defined as per the peak capacity. However, requirement on number of cores for functioning of a RAN NF at medium or low-capacity duration could be much lower compared to the peak capacity. Energy consumed by server/product where RAN NFs are deployed increases with the number of active cores. If any RAN NF is not running at its full capacity, some of the cores can be moved to suspend state and hence can result in reduced energy consumption by server/product. This control can be done at NF level and response time will be at slot level.

Loading conditions change at different nodes at different times, and with different rates. It can change even within one node between different functional entities. Hence decision to move cores into suspend or active state can be done at a different interval at each node. Energy consumption can be optimized on each Network Function individually and at different time granularity levels. For example, the approach to reduce Energy consumption between day / night can be different than the one for a particular hour in the daytime.

Time granularity at O-CU-CP/O-CU-UP can be much higher compared to O-DU [46]. Core requirements in O-CU-CP/O-CU-UP changes with user connection and throughput. Connection establishment takes around a few milliseconds before even data can start. Hence, suspend time and time granularity in these modules can be in terms of a few milliseconds or even seconds. While in O-DU, its operation depends on slot level. Core consumption at Layer 1 module in O-DU changes with each slot hence network EE calculation can be done at slot level in Layer 1 while in Layer 2, it could be in terms of multiple slots. Suspend duration also can be decided accordingly.

AI/ML models can be the key tools to achieve this level of time granularity as the number of parameters to be monitored, increases as we move towards lower time granularity. Initial training can be done statically in the lab. Site data can be collected at a very sparse interval if needed. Contribution for each parameter can be calculated based on the static data, which will be used for core calculation.

As shown in Figure 5-1, AI/ML models can be defined in O-CU-UP, O-CU-CP and O-DU RAN NFs independently to achieve the network EE on each node. Each component of NF can maintain its own model. The cores requirement is determined

at each component to support the call model in a particular condition. As each node has its own set of parameters, that can be fed to its AI/ML model as an input. Based on these parameters, AI/ML model can output the cores requirement by that node at any instance. Interval at which AI/ML can calculate the core requirement and suspend period can be different for each node/component. As AI/ML model is controlled at each node in real time, it can provide better results. For Layer 2, this calculation can be done at every few milliseconds while for Layer 1, it can be calculated on each slot. O-CU-CP/O-CU-UP can have a bigger granularity of calculating the cores required. Based on this calculation, excess cores can be suspended and moved to deep sleep mode. Suspend duration for each NF also can vary for each component/NF. As AI/ML model exists in the same component/NF, data is not required to be sent out to RIC.

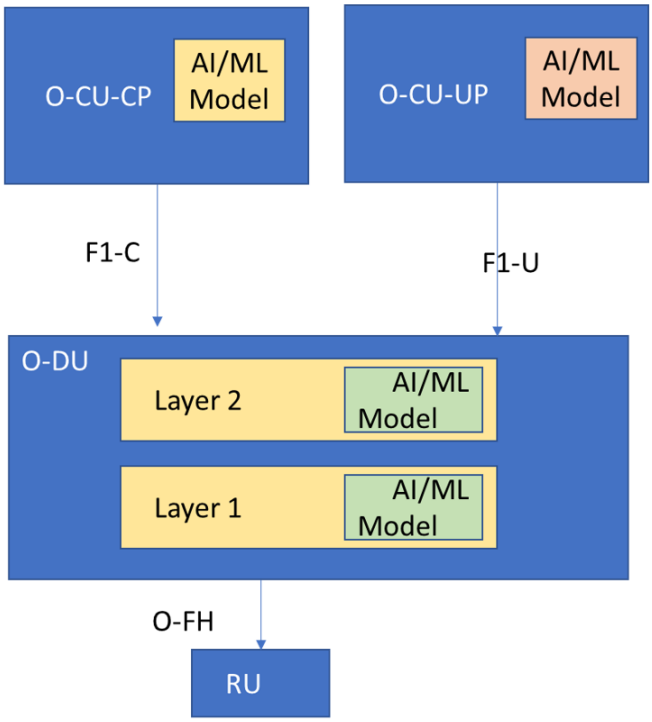

Figure 5-1 AI/ML model in RAN NF [46]

Each NF has different set of parameters impacting its core utilization. Also, there are different functional entities that runs in each node. Each functional entity can consider different parameters for capacity calculation e.g., for O-DU, the following Layer 2 parameters can be considered for capacity calculation:

- Number of Active UEs
- Number of Connected UEs
- Outstanding data at any instance
- Expected throughput etc.

At the same time, for core calculation required to be done at Layer 1, following parameters can impact its core utilization.

- Physical Resource Block (PRB) utilization
- Number of users
- Number of Cells
- Number of layers
- Uplink/Downlink throughput

While for O-CU-CP, it could be:

- Number of connected users
- Number of connected O-DUs
- Number of connected cells, etc.

Above are some of the parameters that can be monitored at their respective nodes/functional entity level. At any point, based on such parameters, number of cores required to run the Layer 2 or Layer 1 in O-DU can be calculated. In low-capacity period, some of the cores can be moved to deep sleep state. By doing profiling, initial weightage of each of the parameter impacting the core utilization can be calculated and it can be used as an input parameter for the AI/ML model to provide cores requirement. Based on this, cores can be kept in active state in O-DU. An AI/ML model can be defined in O-DU to calculate the actual cores requirement for any of the call model running. It can keep tuning the weightage of different parameters to fine tune the system so that better network EE results can be achieved without hampering the system performance.

To cater to cases where the duration of low-capacity conditions is short, cores utilized for running the wireless systems can be reduced. It can be done depending on various factors e.g., in O-DU based on no of Active UEs, Throughput etc. Similarly, even in O-CU-CP/O-CU-UP, few cores can be turned into deep sleep mode to achieve the gain in energy consumption.

6 Network EE in computing and communication

As discussed in previous chapters, the advent of next generation technology promises to improve the way we connect and interact with the digital world. As we move towards a future where the IoT, AR, and AI play increasingly essential roles, it becomes imperative to address the network EE challenges that arise in this evolving landscape. One of the key areas where network EE needs to be a top priority is in the joint computing and communication control of next generation networks. In this chapter we delve into the key features of network EE in the computing and communication domain of next generation networks.

6.1 The growing need for network EE

The proliferation of connected devices, ranging from smartphones to the Internet of Things (IoT) gadgets, and the surge in data-intensive applications such as VR, AR, and ultra-high-definition video streaming, necessitate a substantial increase in computational power and data transfer capabilities. This escalation in demand is pushing the boundaries of current network infrastructure, particularly data centers, which are already energy intensive.

Traditional data centers, which form the backbone of network operations, are known for their high energy consumption. They require substantial power for not just data processing and storage, but also for cooling systems to prevent overheating. As next generation networks are expected to handle significantly more data traffic, this energy consumption is poised to increase dramatically if current trends continue.

To address increased need of power, network EE must be a primary consideration in the design of networks for 6G and beyond. This involves rethinking network architecture, deploying new technologies, and embracing innovative practices. For example, network design innovations such as network function virtualization (NFV) and software-defined networking (SDN) can optimize data routing and reduce redundant data processing, thereby saving energy. Additionally, the integration of AI/ML can further enhance network EE by predicting and managing network loads more effectively.

6G networks also present an opportunity to leverage distributed computing architectures, such as edge computing, which processes data closer to the source. This approach not only reduces the energy consumed in long-distance data transmission but also alleviates the load on centralized data centers.

6.2 Integration of computation and communication

A significant transformation in networks for 6G and beyond is expected on how computation and communication are intertwined. Unlike previous generations, where communication and computation functions were largely separate, next generation aims to merge these two aspects, enhancing both efficiency and performance. This

integration is primarily driven by the adoption of distributed computing paradigms such as edge computing and fog computing [47].

The integration of computation and communication in next generation networks necessitates efficient management of the interplay between these two functions. By combining computational tasks with data processing at the network's edge, the amount of data that needs to be transmitted over long distances is significantly reduced. This not only saves energy but also mitigates the burden on the network's core, which is critical in an environment where the number of connected devices and the volume of data they generate is exponentially increasing.

Furthermore, this integration enables more intelligent and context-aware computing. By processing data closer to where it is generated, the network can make more informed decisions in real-time, enhancing user experience and system efficiency. For example, in a smart city scenario, traffic management systems could process data from various sensors and cameras at the edge of the network, quickly adapting to changing conditions without the need to send all data back to a central server.

6.3 Adaptive resource allocation

In the rapidly evolving domain of telecommunications in the 6G and beyond, the concept of adaptive energy-efficient resource allocation takes center stage, underscoring the necessity for smart and sustainable network management. This approach is centered on the dynamic optimization of resource allocation, encompassing both computational resources and network bandwidth, to align with the real-time demands of a diverse array of applications and devices. This dynamicity is a shift from traditional static resource allocation methods and is vital in a landscape characterized by fluctuating network conditions and diverse user requirements [48].

The cornerstone of this adaptive approach is the integration of sophisticated ML algorithms and AI-driven decision-making mechanisms. These technologies empower the network with the capability to intelligently analyze and interpret vast volumes of data regarding network traffic, user behavior, and the specific needs of various applications. Utilizing the power of predictive analytics, the network can anticipate future resource requirements, enabling proactive adjustments in resource allocation. This predictive capacity ensures that computational power and bandwidth are pre-emptively directed to where they will be most effective, thereby enhancing network efficiency and user experience.

For example, in scenarios where there is an anticipated increase in demand for high-bandwidth services, such as during a major sporting event with numerous users streaming high-definition content, AI algorithms can forecast this increased demand and accordingly allocate greater bandwidth and computational resources to prevent service degradation. Conversely, in periods of lower demand, these resources can be scaled down, conserving energy, and optimizing network operation costs.

The adaptive nature of this resource allocation extends beyond mere reactive adjustments. By integrating AI, the network can learn and evolve its allocation

strategies over time. This learning process enables the network to become increasingly efficient at predicting and managing resource distribution, leading to a continuous improvement in the network EE.

Furthermore, this approach enables the customization of computation and communication resources to cater to the distinct requirements of various applications and services. High-priority or latency-sensitive tasks, such as those essential for autonomous vehicles or critical medical applications, can be assured the requisite resources for uninterrupted operation. Concurrently, less critical applications can have resources allocated in a manner that optimizes overall network efficiency, including network EE, without causing a significant impact on service quality.

Beyond operational efficiency, adaptive energy-efficient resource allocation in next generation networks is pivotal in reducing energy consumption. By intelligently allocating resources according to actual need and scaling resources in response to fluctuating demand, the network significantly lowers its energy usage. This is increasingly important in the context of global environmental challenges and the pressing need for sustainable and green technology solutions.

6.4 Cross-layer optimization

Cross-layer optimization emerges as a pivotal strategy for achieving network EE in the context of next generation networks. This approach transcends traditional networking layers, fostering a coordinated effort between the physical layer, the network layer, and the application layer. The fundamental idea behind cross-layer optimization is to break down the silos that typically exist in network architecture, allowing for more holistic and efficient management of resources across different layers of the network [49].

In a typical network structure, the physical layer is responsible for the actual transmission and reception of signals. The network layer handles the routing of data packets through the network, and the application layer is where end-user applications reside, dictating the type and amount of data that needs to be transmitted. In conventional network architectures, these layers operate largely independently, following predefined protocols and functions. However, this siloed operation often leads to inefficiencies, particularly in terms of energy usage.

Cross-layer energy-efficient optimization seeks to remedy this by ensuring that these layers are not just functioning independently but are actively coordinated and informed by each other's operations. For example, information from the application layer about data priority or required QoS can inform how the physical layer transmits data, choosing energy-efficient transmission methods or adjusting signal strength appropriately. Similarly, insights from the physical layer about current network conditions can help the network layer optimize routing decisions to avoid congested or energy-intensive paths.

Minimizing redundancy is another key aspect of cross-layer optimization. Data processing and transmission consume energy, and unnecessary duplication of these

processes across different layers can lead to significant energy waste. By ensuring that data is processed and routed in the most efficient manner possible, from the point it enters the network to when it reaches its destination, overall energy consumption can be significantly reduced.

Moreover, cross-layer optimization involves smart data routing strategies. Efficient routing algorithms that take into account the energy status of network nodes and the quality of network links can prevent energy wastage. For instance, routing algorithms can be designed to prefer paths that use less energy or avoid nodes that are low on energy, thereby prolonging the lifespan of the network and reducing the overall energy requirement for maintaining the network.

6.5 Green networking technologies

Green networking technologies are essential for the evolution of next generation networks, ensuring a sustainable approach to both performance and environmental concerns. This involves leveraging low-power hardware advancements, incorporating dynamic voltage and frequency scaling mechanisms for optimized EE, and embracing cutting-edge semiconductor technologies. Renewable energy sources, including solar and wind solutions, play a crucial role in reducing dependence on traditional power grids. Implementing smart energy management systems and energy-harvesting technologies ensures continuous network operation while establishing a self-sufficient energy ecosystem. Furthermore, the efficiency of data transfer within next generation networks relies on the adoption of energy-efficient transmission protocols, involving advanced modulation schemes, error correction techniques, and signal processing algorithms to achieve higher data rates with reduced power consumption. This comprehensive integration of green networking technologies positions 6G and beyond, as a technologically advanced and environmentally responsible telecommunications paradigm [50].

It is notable that the disaggregated nature of O-RAN aligns with green networking principles by allowing for the selection of energy-efficient hardware components from different vendors. We predict that O-RAN will be able to complement the deployment of renewable energy sources across distributed network components and consequently reduce dependency on traditional power grids [51]. This flexibility is key in integrating renewable energy sources like solar, wind, or other sustainable power options directly into the network infrastructure. In a traditional, monolithic network setup, the components are typically standardized and often less adaptable to specific energy-saving technologies. However, with O-RAN, operators can specifically select and configure components that are designed to work efficiently with renewable energy sources. For instance, BSs and other network elements can be powered by local renewable energy systems, reducing the need for electricity from conventional power grids.

Moreover, the modular nature of O-RAN means that network elements can be distributed across various locations, which allows for the strategic placement of renewable energy systems. This distribution not only enhances the overall energy

efficiency but also ensures that energy needs are met locally, decreasing reliance on centralized, traditional power supplies.

Additionally, O-RAN's advanced software and control mechanisms can dynamically manage energy consumption, further optimizing the use of available renewable energy and minimizing the draw from traditional grids. This intelligent energy management aligns with green networking principles and supports the broader goal of reducing carbon footprints and enhancing sustainability within telecommunications networks.

There are ample opportunities for workload acceleration in both RAN and core network functions, including acceleration of signal processing algorithms in physical layer, scheduling algorithms in medium access control layer, AI-aided beamforming techniques in the RAN domain, and acceleration of various user plane functions in the core domain such as general packet radio service (GPRS) tunneling protocol, packet classification, receive side scaling (RSS) hashing etc. Striking the right balance between performance and EE trade-off can be achieved by accelerating full stack of network functions (especially the compute intensive workloads in each layer) which can improve the network performance under the same power envelope, resulting in better EE of the overall network [52]

6.6 Energy aware design standards

Next generation standards bodies like O-RAN ALLIANCE and industry stakeholders must establish EE as a core design principle. Developing energy-aware design standards will drive innovation and create a framework for building sustainable next generation networks that prioritize energy conservation without compromising performance.

7 Construction strategy for an energy-efficient AI/ML platform

AI/ML technologies are at the forefront of modern computing, driving innovations across numerous industries. However, these breakthroughs are considered as energy-hungry technology. Hence, the energy consumption of these technologies has become a pressing concern. Since the adoption of AI/ML technology is massive and widespread, thus, an energy-efficient AI/ML platform is necessary to be implemented. Therefore, this chapter focuses on the construction strategies necessary for developing an energy-efficient AI/ML platform, ensuring sustainability, and without decreasing performance.

7.1 Sustainable design methodology for AI/ML platform

As the sustainable development goal is a new demand for entities, either academic or industry, to protect the environment from degradation and to avoid the deterioration of available resources, therefore, the adoption of sustainable development schemes or green infrastructure becomes a trend that needs to be fulfilled. Furthermore, a new paradigm is necessarily triggered by massively endorsing sustainable approach and investing in green architecture.

Creating a sustainable AI/ML platform starts with an appropriate assessment of the implemented technique and a robust design methodology. This approach encompasses not only the efficient use of energy but also considers green concept or paradigm into the architecture of the platform. AI/ML platforms can significantly reduce the production of the carbon footprint while maintaining high performance [53], [54].

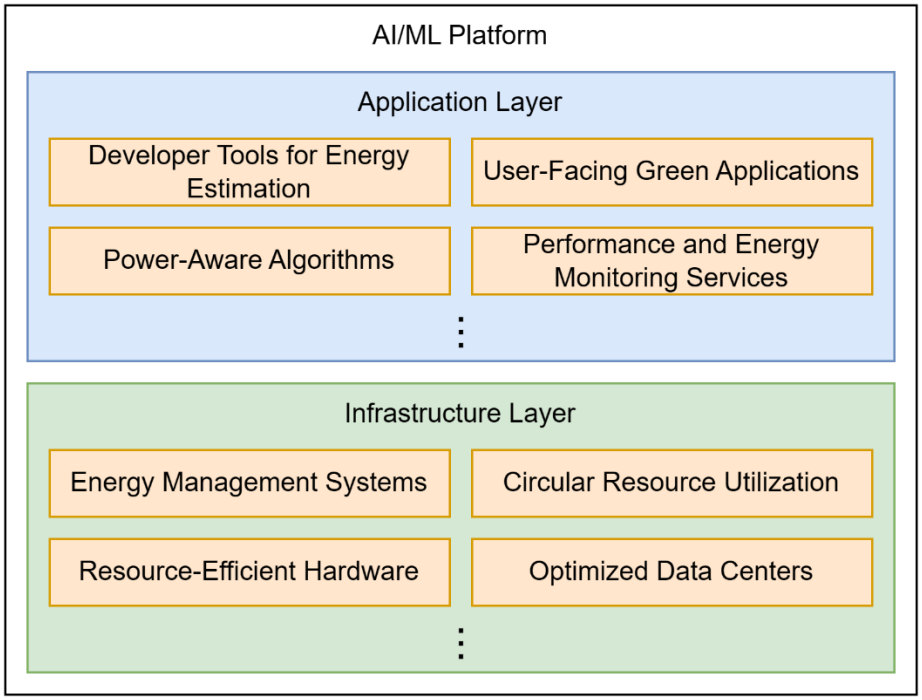

Figure 7-1 Sustainable AI/ML platform with the infrastructure and the application layer

In the architecture design for AI/ML systems, as shown in Figure 7-1, it is crucial to incorporate a layer focused on developer tools and another on application services built upon the foundational computing infrastructure. This underlying infrastructure involves managing multiple servers, which inherently includes addressing power supply needs. To support effective decision-making for future power grids, it is essential to integrate a service at the developer-focused function that predicts power needs of processes. Additionally, at the application layer, a service should provide estimations of power loss during process execution. Such estimations enable us to determine the total power consumption cost for training models and optimize energy distribution. This structured approach not only boosts the EE of the AI/ML platform but also enables more precise control over distributed energy resources, boosting the overall energy effectiveness.

Sustainable development infrastructure is established by creating a circular utilization of each resource. Circular utilization on the infrastructure is the concept to utilize the idle resources and wisely construct the infrastructure as needed without producing excessive wasteful resources. Therefore, by incorporating EE and circular utilization concepts into the design and development of AI/ML platforms, the sustainable AI/ML platform is not only applicable, but it also can be a solution to constructing a green infrastructure. By applying various optimization strategies across different levels of the technology stack—from foundational infrastructure to developer tools and user-facing applications, developers of each technology stack can have an estimation of the energy, prevent energy loss, and initiate to form an innovative sustainable development methodology [55], [56]. This methodology not only helps us minimize energy consumption and carbon footprint while improving the performance of AI/ML platforms, but also provides an advanced paradigm and approach for achieving sustainable development goals [55].

Highly efficient utilization of the resources plays a significant key in the reduction of the carbon footprint. As mentioned, optimization strategies to have efficient materials to construct the AI/ML platform could reduce the consumption of electric power as well as avoid excessive stuff being put into the platform. Therefore, there are no or, at least, fewer idle resources inside the platform. Optimizing resource use within the platform minimizes idle resources, ensuring that every component is effectively utilized. This approach not only conserves energy but also enhances the platform's overall efficiency in resource management.

7.2 Energy-efficient microservice design and architecture

In recent years, there has been an increasing demand for AI/ML platforms that are not only powerful and scalable but also energy-efficient. As organizations continue to deploy these platforms at scale, the energy consumption associated with running complex computations and maintaining large-scale systems has become a significant concern. Microservices architecture, known for its ability to decompose applications into smaller, loosely coupled services, plays a crucial role in addressing these challenges. By designing these microservices to be energy-efficient, organizations can significantly reduce the overall energy footprint of their AI/ML platforms. The exploration of best practices in creating scalable architectures that are not only

efficient in processing but also in energy usage is crucial to ensure the scalability of the maintenance and deployment of the platform.

Each microservice within the pipeline is designed to be highly modular, focusing on a specific task such as data preprocessing, feature engineering, model training, model evaluation, as well as model deployment [57]. This fine-grained decomposition allows for microservices to be used "as needed", reducing unnecessary computation and thus conserving energy [58]. By enabling dynamic scaling, these microservices can adjust their resource consumption based on real-time demands, which not only enhances EE but also ensures optimal performance under varying load conditions. Additionally, leveraging state-of-the-art techniques in load balancing and energy-efficient hardware can significantly reduce the power consumption of these microservices.

From a practical perspective, a pipeline's workflow typically consists of multiple microservices. The traditional approach is to execute only one stage at a given point in time. To optimize performance, we can establish a service that simultaneously analyzes the optimal execution strategy for the pipeline. Additionally, we need to consider how to deploy these microservices to high-efficiency nodes and how to design the cluster [59].

In terms of deployment strategy, we can optimize the design based on the geographical location of the servers to avoid excessive network transmission nodes, thereby reducing energy consumption. To assess energy consumption more precisely, we can introduce the concept of calculating energy consumption on a per-microservice basis within the nodes. By calculating the energy consumption caused by data transmission between different nodes for each microservice, we can select the optimal deployment plan and allocate microservices to nodes with minimal energy consumption [60], [61].

By adopting an efficient microservice design, the network can minimize losses and maximize resource utilization, leading to a more effective allocation of the available infrastructure. This approach not only reduces energy consumption but also allows for cost savings, contributing to the development of a robust, compact, and eco-friendly AI/ML infrastructure. Additionally, using optimized and scalable microservices reduces the need for intensive monitoring of each service within the platform, enabling smarter and more effective allocation of available resources.

Besides, the deployment of shared microservices can also be a considered methodology to provide optimal solutions and efficiency on the deployment of the services. For instance, as illustrated in [62], one microservice for implementing the sustainable Internet of Vehicles (IoV) can serve several tasks. By applying correct scheduling and optimization on the time-varying network load, the resource utilization and service latency can be guaranteed. Therefore, the scalability of the services can be maintained and controlled. Eventually, once the additional tasks are deployed in the microservices, the growth of the resources will not be exacerbated and excessive.

7.3 High-performance big data architecture using heterogeneous cloud

To address the growing demand for high-performance and energy-efficient AI/ML platforms, it's crucial to explore advanced cloud technologies. Many organizations seek scalable solutions that not only meet their computational needs but also optimize energy consumption. This is especially relevant in environments where data volumes and computation demands are constantly increasing, making traditional single-cloud solutions insufficient.

Leveraging heterogeneous cloud environments is a highly effective approach for achieving a balance between performance and EE in an AI/ML platform. This chapter explores the utilization of diverse cloud services to optimize data processing tasks, focusing on high performance with minimized energy consumption. Through the implementation of Software-Defined Networking (SDN), combined with intelligent services, the platform can establish efficient data transmission mechanisms. This enables the selection of optimal network paths, effectively integrating Edge-cloud and Multi-access Edge Computing (MEC) with traditional cloud services. Such a strategy enhances the overall efficiency of data processing and routing, contributing to better performance and lower energy use across different cloud architectures [61], [63].

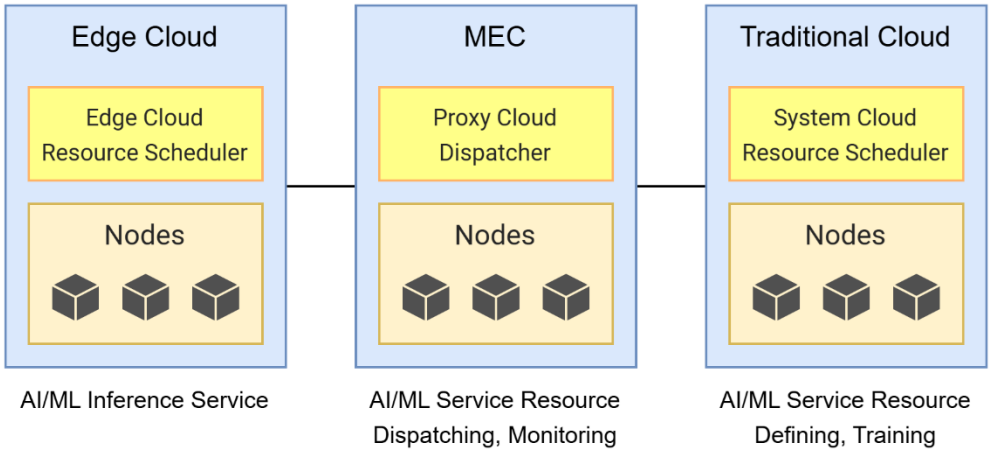

Figure 7-2 Three-tier resource scheduling architecture in heterogeneous cloud

By segmenting heterogeneous cloud services into a three-tier resource scheduling architecture as shown in Figure 7-2, effective resource utilization can be achieved. In network provisioning services, a large amount of heterogeneous information is generated. In traditional heterogeneous cloud services, this diversity in information often leads to uneven resource utilization and high manual maintenance costs, which has impacted the operational efficiency of network providers in past cases [64].

Mainly, heterogeneous cloud services are established by adopting a three-tier resource scheduling architecture, comprising a system cloud resource scheduler, a proxy cloud dispatcher, and an edge cloud resource scheduler. The system cloud resource scheduler is responsible for defining and regulating AI/ML service resources at the initial stage and adjusting the AI/ML model architecture. The proxy cloud

dispatcher efficiently dispatches the fully defined AI/ML service resources to the edge cloud resource scheduler, which monitors the AI/ML inference information. Ultimately, the edge cloud resource scheduler controls the Infrastructure as a Service (IaaS) tool management system, completing the scheduling of the heterogeneous network system from public to private clouds [65].

The deployment on the heterogeneous cloud is suggested to accommodate the characteristics of each environment. First, in edge cloud, due to relatively limited computational resources, the optimized and lightweight models are appropriate to be employed. In the deployment point of view, the optimized and light models have quick and fast response especially for inference on the edge side. On the edge cloud, the computation using AI/ML models can be performed regardless of the support on the advanced hardware such as TPU and GPU as long as the model is optimized enough. Hence, GPU or TPU support may accelerate the AI/ML models running on the edge cloud. For the utilization of the cloud computing platforms, the adjustment of the cloud cluster usage is simply performed by calling the API for the provider. The orchestration of this edge-cloud scheme then has an impact on the separation of the training and inference task. Mostly, training tasks are executed on the cloud and inference is conducted on the edge. Therefore, using this architecture, the network loss, which refers to the loss of data packets during transmission due to factors such as signal attenuation, interference, congestion, and latency, can be minimized while reducing the efforts for transmitting redundant and useless data package to the cloud.

Thus, from the data transmission point of view, performing real-time inference using AI/ML models is conducted on edge devices which can ensure the low latency computation and quickly generate the inference result. Moreover, data is stored in the cloud which is also used to train AI/ML models while the QoS settings can be adjusted to maximize the cooperation between edge-cloud and MEC.

Through this design, we can fully utilize the advantages of heterogeneous cloud environments, improve overall processing performance, and reduce energy consumption at the same time [66].

7.4 Methodologies for energy-efficient Inference in AI/ML model

Energy-efficient inference in AI/ML models is crucial for reducing operational and investment costs such as the investment on the platform construction, the cost on deployment, and maintenance costs. The scheme to optimize the inference model is essential to reduce energy consumption. Thus, the optimization schemes of inference algorithms are elaborated in this chapter specifically to introduce relevant concepts and methods.

Achieving energy-efficient inference requires efforts from multiple aspects, including the selection of framework, model, and hardware. In addition, to reduce the computational energy during training and inference stage, it is appropriate to choose the necessary model. For instance, instead of trying to build the complex and large AI/ML model to be deployed, we can consider employing the model which has less complexity yet can fulfill the requirement of the task.

Selecting suitable frameworks and models can also reduce computational complexity and decrease energy consumption. Utilizing hardware designed specifically for AI/ML tasks, such as TPUs and GPUs, can further improve computational efficiency and lower energy usage. However, employing TPUs and GPUs also needs to consider the usage and the deployment scenario. Wisely selecting the hardware being used for the deployment is also essential to reach the optimal solution.

Currently, common model optimization techniques include pruning and knowledge distillation. Pruning reduces model size and computational complexity while maintaining model performance by removing redundant or unimportant parts of the model. Knowledge distillation is a compression method based on teacher-student models, where knowledge from a large, complex model is transferred to a smaller model, resulting in a compact model with performance comparable to the large model, thus reducing computational cost and energy consumption during inference [67]. Both pruning and knowledge distillation can reduce the number of deployed parameters. By minimizing the number of parameters, we can avoid excessive resource allocation while the impact of this optimization is not only saving the energy and carbon being released to the earth, but we can also save the cost of operations and investment.

Quantization is another common technique for model compression, where trained model weights are quantized to decrease the precision of decimal points for each weight. For example, floating-point numbers are converted to integers, leading to a decrease in overall parameter size while still maintaining a certain level of inference accuracy. This achieves the goal of improving inference efficiency by striking a balance between inference accuracy and efficiency [68].

Quantization plays a vital role for real-time application especially for the implementation of the AI/ML models in the edge computing side. Commonly, the hardware resources in edge computing are less powerful compared to the cloud one. In the edge side, a Reduced Instruction Set Computer (RISC) architecture of the Advanced RISC Machine (ARM) based CPU is often selected to be deployed. Therefore, optimization using quantization is unavoidable to fully exploit the computational resources and ensure to meet the real-time performance.

Additionally, the developer and algorithm designers may also consider in combining the pruning, knowledge distillation, and quantization strategy. By training the model using pruning and knowledge distillation, it is known that we can maintain the performance of the model [69]. Furthermore, before the deployment of the trained model, quantization can be applied to compress the computational complexity and accelerate the inference time.

By judiciously selecting frameworks, models, and hardware, and applying appropriate model optimization techniques such as pruning, knowledge distillation, and quantization, the inference efficiency of AI/ML models can be significantly improved, reducing energy consumption, lowering operational costs, and promoting the sustainable development in real-world scenarios [70].

8 EE in 6G and beyond telco cloud and edge data centers

8.1 Introduction

Telco cloud and edge uses cloud computing technology along with NFV and SDN technologies to virtualize and containerize the telco network functions, i.e., in forms of Virtual Network Functions (VNFs) or Cloud-Native Network Functions (CNFs), so that operators can manage their infrastructure and network resources efficiently. It gives an opportunity to operators to address the issues of dynamic resource demands from applications and establish open platforms for providing new services [71] [72]. In other words, Telco cloud provides a cloud-based environment optimized for the workloads and requirements of telecom operators [73].

EE is currently considered as one of the main research directions for 6G [74]. Telco Cloud datacenter has played an important role on energy consumption of the mobile network. One research effort showed that around 9% of energy consumption of the existing mobile network is related to its Telco Cloud & Edge datacenter [75], [76]. Improving EE in data centers has attracted a lot of attention in recent years because of its high impact on economy (e.g., energy cost), environment (e.g., carbon emission), and system performance (e.g., system performance-energy savings relation) [77]. Looking at the main 6G objectives, it is expected that 6G offers a noticeable EE over the 5G network [78]. It makes the topic of EE for Telco Cloud & Edge datacenter more important than ever before.

In general, the data center industry has seen success in driving EE in centralized and regional data centers. While a standard data center had a Power Usage Effectiveness (PUE) of around 2 about 20 years ago, today we can find some data centers with a PUE of around 1.5 [79] and even less than 1.1 [80]. But, looking at 2030 perspective or recent mandatory reporting for data centers [81], it is still necessary to work on improving EE of data centers. For example, the European Commission defined the following key action in the Shaping Europe's Digital Future document [82]: "initiatives to achieve climate-neutral, highly energy-efficient and sustainable data centers by no later than 2030 and transparency measures for telecoms operators on their environmental footprint". As another example, according to the EE Act (Energieeffizienzgesetz, EnEfG), data centers with operations starting on or after July 1, 2026, must achieve PUE of less than or equal to 1.2 in Germany [83]. Notice that the ideal PUE is 1.0, where all the energy required by the data center would go directly to powering its IT equipment. However, the need for cooling, lighting etc. makes reaching values close to 1.0 extremely difficult [84]. A PUE of 2.0 means that for every watt of IT power, an additional watt is consumed to cool and distribute power to the IT equipment [81].

There are currently many solutions for improving EE of data center ranging from design, and cooling to data center management. We introduce some current and potential techniques and technologies that can be used for improving EE in Cloud data centers. Before introducing them, we present the concept of next generation wide-

area cloud and highlight the main difference between 6G and 5G in the aspect of computing.

8.2 6G wide-area cloud

5G currently supports modern compute-intensive applications and technologies such as VR, AR, metaverse, and AI. In development for 2030, it is expected 6G will provide more support (e.g., from EE and QoS aspects) for modern compute-intensive applications and technologies. To this end, one potential solution under discussion is called 6G wide-area cloud. The 6G cloud-native system enables a wide-area cloud unifying the computing functionality of mobile devices, mobile networks, and datacenters. This 6G feature allows computing resources from mobile devices, mobile networks, and data centers to be coherently shared in the form of a wide-area distributed cloud, and jointly provide computing services to user applications. In other words, the 6G system will be built on the cloud and can natively provide computing services (i.e., computing becomes part of the system service). It will be different from 5G, where clouds (either edge, regional or central clouds) reside in the Data Network (DN) beyond mobile core network. Nevertheless, the 6G system is expected to be revolutionary (*i.e*., different from 5G in some respects) and evolutionary (*i.e.,* an extended version of 5G in some aspects). Therefore, 5G edge computing and 6G wide-area cloud still have common features. One common feature is that they both use computing resources, with 6G further pushing the distribution of computing [85]. Putting all pieces together, it is expected 6G utilizes different sizes of energy-efficient data centers ranging from large-size data centers to a high number of distributed smaller micro and nano data centers at the edge.

8.3 Current and potential strategies for energy-efficient 6G Cloud data centers

Datacenter life cycle is composed of six fundamental phases: plan, design, build, commission, operate and maintenance, and assess [86], [87]. Planning, design, and operation management are the most important phases in aspect of EE [88]. Here, we introduce several existing and potential technologies and techniques that can be used for improving EE of 6G Cloud datacenters. These solutions are mainly utilized in planning, design, and operation management phases.

Additive manufacturing technology: the process of fabricating three-dimensional objects by depositing materials layer-by-layer directly from computational geometry model is called additive manufacturing. It eliminates the design and fabrication limitations of conventional manufacturing approaches to a large extent [89]. Additive manufacturing provides more flexibility to customize the design, and then enables really complex designs that cannot be realized with traditional manufacturing [90]. As example, additive manufacturing technology has been used to design and build an integrated, leak-free unibody liquid cooled heat sink. It could reduce PUE of a datacenter [90], [91].

Modular data centers: Modular data centers include all the resources that are required to build out data center space in an integrated package. They are self-

contained units (i.e., in a variety of shapes and sizes). As already mentioned, it is expected 6G utilizes a high number of small edge data centers, almost everywhere. One benefit of modular data centers is that it can offer flexibility by starting with small installations and increasing them in size based on need. For large-size data centers, while a conventional data center can take about two years to be installed from conceptualization to deployment into functional use, it can be much faster for a modular data center, often taking 50 to 75% less time. Last but not least, cooling efficiency is another feature of modular data centers. According to Building Industry Consulting Service International (BICSI), modular data centers can be up to 40% more energy-efficient than an open data center environment [92].

Digital Twin AI: AI-based mechanisms including ML are expected to play an important role in improving EE of Cloud data centers. For example, more accurate ML-based prediction mechanisms can be designed to provide more insights into the future demand of a particular resource (e.g., CPU) based on historical workload. These predictions can be utilized to deal with non-linear resource usage and energy consumption in Cloud data centers [93]. Secondly, DT can be used as a tool to provide a digital representation of Telco Cloud datacenters. The DT model can be utilized to optimize AI strategies. It can provide extra data sets for training and solve the problem of insufficient data in a real data center. For example, a combination of AI and DT can offer promising solutions to improve EE of data centers by optimizing air distribution and cooling redundancy [94].

Dynamic rightsizing and consolidation: An old but still efficient way for reducing energy consumption of data centers is to dynamically adapt the number of active servers to match the current workload (i.e., to dynamically “right-size” or scaling the data center). In other words, dynamic right-sizing means adapting the way requests are dispatched to servers in the data center in a way that, during low load time, servers that do not have jobs enter a power-saving mode (e.g., go to sleep or shut down) [95].

On the other hand, consolidation technique is utilized to consolidate multiple Virtual Machines (VMs) or containers onto a reduced number of servers. By running fewer physical servers (i.e., right-sizing), energy consumption of data centers can be reduced. Although there are already many solutions for consolidation, by emerging new cloud architectures, data centers and technologies, new energy-efficient resource allocation and consolidation methods are still required.

Network functions consolidation: In a typical cloud deployment, diverse workloads like virtual RAN (vRAN) network functions, distributed User Plane Function, and security functions (e.g., firewall) can run in the same cloud datacenter, and even on the same server rack. Underutilized server platforms lead to higher energy consumption [52]. Therefore, it is suboptimal for a cloud datacenter to use many low-utilization servers compared to a compact accelerated datacenter with a few optimally utilized accelerated computing servers. One way to achieve this consolidated hosting of network functions/applications and efficient processing is to utilize a fully software-defined cloud infrastructure deployed on general-purpose hardware platforms, enabling acceleration of key network functions. Besides the latency benefits of network functions’ co-location, densely packing network functions on a homogeneous cloud

platform result in a higher resource utilization and overall datacenter energy savings. A densely packed and software-defined accelerated datacenter built upon CPU, GPU, and DPU-based architecture on commercial-off-the-shelf (COTS) servers can support consolidation of network functions in the RAN and 5G Core for vRAN workload, and at the same time, can also host other non-vRAN workloads like emerging generative AI and large language model (LLM) based applications, which will open up new business opportunities for telco cloud data centers.

Shared AI and telecommunications infrastructure: The advent of generative AI will significantly influence the landscape of 6G. The telecommunication industry is already seeing unparalleled rise in compute demand for executing plethora of LLM based applications at the edge, driving massive growth of edge GPU-based servers for running LLM inferencing and other related multi-access edge compute (MEC) applications. The proliferation of heterogeneous applications running on the same network infrastructure is fueling more and more isolated clusters within cloud data center dedicated for running multitudes of workloads and diverse applications of the network. Not all these isolated clusters' hardware resources are fully utilized all the time due to dynamic nature of network traffic density, resulting in significant underutilization of compute resources throughout the network cloud infrastructure, and as a direct consequence, an overall increase in cloud data center's carbon footprint. One way to resolve this EE suboptimality is to build multi-tenant cloud infrastructure that is software-defined and can run heterogeneous workload (like RAN, Core, AI and MEC applications) by sharing the same compute resources of the underlying cloud platform. Agile and intelligent orchestration of AI and network functions workload on the same cloud data center infrastructure can lead to optimum compute resource utilization. The software-defined resources can be repurposed and shared across heterogeneous workload dynamically over time, as per the instantaneous compute needs, leading to resource overhead reduction and overall improved EE of network operation.

Energy-efficient quantum data centers: Quantum computers capabilities will be much larger than current systems. Therefore, it is expected we have Quantum Data Centers. However, as most of the proposed quantum computing systems need extremely low temperatures (mK to K), quantum data centers should have very different thermal management approaches from conventional data centers. Therefore, the cooling system not only needs to remove the heat entering the cryostat from the ambient environment, but it also needs to eliminate all power dissipated by electronics operating at low temperatures from a cryostat. Because of this, the power consumption of the cooling system is likely to be much greater than the power consumption of the electronics. In other words, quantum systems utilize energy in very different ways than conventional computers. Instead of the energy use being dominated by energy used within the electronic circuitry, energy use is dominated by cooling requirements. Therefore, finding solution for improving EE of quantum data centers is considered as one of the future research topics [96].

9 Conclusion

The comprehensive analysis conducted in this research report sheds light on the paramount importance of network EE in the evolution of next generation networks. As the telecommunications industry advances towards the 6G and beyond era, it becomes increasingly crucial to embed EE as a foundational principle in the design and operation of network infrastructures. The integration of Green Communication principles into next generation networks is not merely a technical upgrade but a strategic necessity to ensure sustainable and environmentally responsible network operations. The evolution from 5G to 6G emphasizes the need for advanced network EE metrics and integrating AI/ML to optimize energy consumption across various network functions. The integration of AI-driven methodologies for network management is crucial for achieving sustainable and energy-efficient next generation networks.

This research explores strategies for enhancing network EE, focusing on implementing energy saving mechanisms, improving hardware, optimizing radio planning, centralizing processing, and employing intelligent energy management in cloud computing, all geared towards sustainability and innovation.

It further explores the potential of AI/ML models to enhance network EE, focusing on dynamic resource allocation and intelligent network management. These models are instrumental in optimizing the energy usage of network functions, thereby reducing the overall energy consumption of the network.

The report delves into the computational and communicative aspects of network EE, highlighting the integration of edge computing to reduce energy consumption and the role of AI models in managing network loads and resources efficiently. It emphasizes adaptive resource allocation and cross-layer optimization as essential strategies for enhancing network EE. Moreover, the adoption of green networking technologies and energy-aware design standards is advocated to reduce the environmental impact and improve the sustainability of next generation networks.

The deployment of an energy-efficient AI/ML platform requires integrating sustainable design methodologies with resource management strategies. By focusing on energy optimization across the entire system architecture from core infrastructure to application-level services, these platforms can effectively reduce energy consumption while ensuring high performance. Key strategies include dynamic resource allocation, intelligent network management, and microservices optimization, all crucial for achieving an optimal balance between EE and operational effectiveness.

In the context of telco cloud and edge data centers, the report identifies their significant impact on the network's energy consumption, advocating for the design of energy-efficient data centers to support the next generation infrastructure. The adoption of innovative technologies and strategies for EE in data centers is highlighted as a key factor in achieving the broader goal of an energy-efficient and sustainable next generation ecosystem.

In conclusion, the report articulates the necessity of incorporating Green Communication principles into the design and operational framework of next generation networks. It proposes that a strategic shift towards energy-efficient technologies and methodologies is essential to realize the vision of sustainable and environmentally friendly next generation communication systems. The commitment to EE not only addresses the technological advancements required for next generation networks, but also aligns with the global imperative for environmental sustainability in the telecommunications sector.

References

[1] ASG Andrae, T. Edler, On Global Electricity Usage of Communication Technology: Trends to 2030. Challenges, 2015.

[2] ITU-R M.2516-0: Report ITU-R M.2516-0 (11/2022)

[3] 3GPP TS 28.554: "3rd Generation Partnership Project; Technical Specification Group Services and System Aspects; Management and orchestration; 5G end to end Key Performance Indicators (KPI)", Release 18, December 2023

[4] KEPLER: Kepler, sustainable-computing.io

[5] K. Yang, S. Jin, N. Rajatheva, J. Hu and J. Zhang, "Energy self-sustainability in 6G," in *China Communications*, vol. 17, no. 12, pp. iii-v, Dec. 2020

[6] W. Wu, C. -S. Yang, I. -K. Fu, P. -K. Liao, D. Calin and M. Fan, "Revisiting the System Energy Footprint and Power Efficiency on the Way to Sustainable 6G Systems," in *IEEE Wireless Communications*, vol. 29, no. 6, pp. 6-8, December 2022.

[7] What is 5G Energy Consumption? https://www.viavisolutions.com/en-us/what-5g-energy-consumption

[8] A technical look at 5G energy consumption and performance, https://www.ericsson.com/en/blog/2019/9/energy-consumption-5g-nr

[9] Parsing the 5G power equation: Is 5G actually greener? https://www.rcrwireless.com/20220124/5g/parsing-the-5g-power-equation-is-5g-actually-greener

[10] B. Mao, F. Tang, Y. Kawamoto and N. Kato, "AI Models for Green Communications Towards 6G," in *IEEE Communications Surveys & Tutorials*, vol. 24, no. 1, pp. 210-247, 2022

[11] 6G networks will be energy efficient from the get-go thanks to AI/ML, https://www.bell-labs.com/institute/blog/6g-networks-will-be-energy-efficient-from-the-get-go-thanks-to-aiml/

[12] H. M. F. Noman *et al.*, "Machine Learning Empowered Emerging Wireless Networks in 6G: Recent Advancements, Challenges and Future Trends," in *IEEE Access*, vol. 11, pp. 83017-83051, 2023.

[13] X. Huang, K. Zhang, F. Wu and S. Leng, "Collaborative Machine Learning for Energy-Efficient Edge Networks in 6G," in *IEEE Network*, vol. 35, no. 6, pp. 12-19, November/December 2021.

[14] W. Y. B. Lim, N. C. Luong, D. T.Hoang, Y. Jiao, Y. C. Liang, Q. Yang, B. T. Kang, "Federated learning in mobile edge networks: A comprehensive survey". *IEEE Communications Surveys & Tutorials*, 22(3), 2031-2063. (2020).

[15] A. Kaul and J. Gupta, "Revolutionary 6G: Technologies, Architecture, Coverage, and Performance," *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2021.

[16] K. Yang, S. Jin, N. Rajatheva, J. Hu and J. Zhang, "Energy self-sustainability in 6G," in *China Communications*, vol. 17, no. 12, pp. iii-v, Dec. 2020

[17] A Gharehgoli *et al*. "AI-based Robust Resource Allocation in End-to-End Network Slicing under Demand and CSI Uncertainties." *arXiv preprint arXiv:2202.05131*. 2022.

[18] H Hojatian*, et al*. "Learning Energy-Efficient Hardware Configurations for Massive MIMO Beamforming." *arXiv preprint arXiv:2308.06376* (2023).

[19] U. M. Malik, M. A. Javed, S. Zeadally and S. u. Islam, "Energy-Efficient Fog Computing for 6G-Enabled Massive IoT: Recent Trends and Future Opportunities," in *IEEE Internet of Things Journal*, vol. 9, no. 16, pp. 14572-14594, 15 Aug.15, 2022.

[20] O. L. A. López, H. Alves, R. D. Souza, S. Montejo-Sánchez, E. M. G. Fernández and M. Latva-Aho, "Massive Wireless Energy Transfer: Enabling Sustainable IoT Toward 6G Era," in *IEEE Internet of Things Journal*, vol. 8, no. 11, pp. 8816-8835, 1 June1, 2021.

[21] A. Alawadhi and A. Almogahed, "Recent Advances in Edge Computing for 6G," *2022 International Conference on Intelligent Technology, System and Service for Internet of Everything (ITSS-IoE)*, Hadhramaut, Yemen, 2022, pp. 1-6

[22] X. Yang, Z. Zho and B. Huang, "URLLC Key Technologies and Standardization for 6G Power Internet of Things," in *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 52-59, June 2021.

[23] C. Sun, C. She and C. Yang, "Energy-Efficient Resource Allocation for Ultra-Reliable and Low-Latency Communications," *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Singapore, 2017, pp. 1-6,

[24] C. She *et al.*, "Deep Learning for Ultra-Reliable and Low-Latency Communications in 6G Networks," in *IEEE Network*, vol. 34, no. 5, pp. 219-225, September/October 2020,

[25] K. S. Reddy, S. Bhagwath, N. Ragavenderan, V. K, G. N. Naik and N. G. S, "Traffic Data Analysis and Forecasting," *2023 7th International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, Bangalore, India, 2023, pp. 1-6

[26] Alcaraz López, Onel L., et al. "Massive Wireless Energy Transfer: Enabling Sustainable IoT Towards 6G Era." *arXiv e-prints* (2019): arXiv-1912.

[27] K. Yang, S. Jin, N. Rajatheva, J. Hu and J. Zhang, "Energy self-sustainability in 6G," in *China Communications*, vol. 17, no. 12, pp. iii-v, Dec. 2020

[28] C. -X. Wang, J. Wang, S. Hu, Z. H. Jiang, J. Tao and F. Yan, "Key Technologies in 6G Terahertz Wireless Communication Systems: A Survey," in *IEEE Vehicular Technology Magazine*, vol. 16, no. 4, pp. 27-37, Dec. 2021

[29] J. Tan and L. Dai, "THz Precoding for 6G: Challenges, Solutions, and Opportunities," in *IEEE Wireless Communications*, vol. 30, no. 4, pp. 132-138, August 2023

[30] K. Yang, S. Jin, N. Rajatheva, J. Hu and J. Zhang, "Energy self-sustainability in 6G," in *China Communications*, vol. 17, no. 12, pp. iii-v, Dec. 2020

[31] ITU-R, “Framework and overall objectives of the future development of IMT for 2030 and beyond,” International Telecommunication Union Radiocommunication Sector, Recommendation ITU-R M.2160-0, Dec. 2023.

[32] https://hexa-x.eu/wp-content/uploads/2023/01/IMT-20306GPromotionGroup-QingyangWang-Workshop-Jan-2023.pdf

[33] T. Yu, S. Zhang, X. Chen and X. Wang, "A Novel Energy Efficiency Metric for Next-Generation Green Wireless Communication Network Design," in *IEEE Internet of Things Journal*, vol. 10, no. 2, pp. 1746-1760, 15 Jan.15, 2023

[34] X. Chen, X. Wang, B. Yi, Q. He and M. Huang, "Deep Learning-Based Traffic Prediction for Energy Efficiency Optimization in Software-Defined Networking," in *IEEE Systems Journal*, vol. 15, no. 4, pp. 5583-5594, Dec. 2021

[35] F. Han, Z. Safar and K. J. R. Liu, "Energy-Efficient Base-Station Cooperative Operation with Guaranteed QoS," in *IEEE Transactions on Communications*, vol. 61, no. 8, pp. 3505-3517, August 2013

[36] V. Zhang, M. Erol-Kantarci, W. Sun, Y. Dai, J. Hoydis and M. C. Gursoy, "Guest Editorial: AI and 6G Convergence: An Energy Efficiency Perspective," in *IEEE Network*, vol. 35, no. 6, pp. 10-11, November/December 2021

[37] Y. Liang and P. K. Saha, "Energy Consumption Forecasting Based on Long Short-term Memory Neural Network with Realistic Smart Meter Data," *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, Singapore, Singapore, 2022, pp. 1374-1379

[38] M. Shehab, T. Khattab, M. Kucukvar and D. Trinchero, "The Role of 5G/6G Networks in Building Sustainable and Energy-Efficient Smart Cities," *2022 IEEE 7th International Energy Conference (ENERGYCON)*, Riga, Latvia, 2022, pp. 1-7

[39] Rama Krishna, K. Rathor, J. Ranga, A. Soni, S. D and A. K. N, "Artificial Intelligence Integrated with Big Data Analytics for Enhanced Marketing," *2023 International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 1073-1077

[40] Z. Liu, Z. Zhu, J. Gao and C. Xu, "Forecast Methods for Time Series Data: A Survey," in *IEEE Access*, vol. 9, pp. 91896-91912, 2021

[41] C. K. Metallidou, K. E. Psannis and E. A. Egyptiadou, "Energy Efficiency in Smart Buildings: IoT Approaches," in *IEEE Access*, vol. 8, pp. 63679-63699, 2020

[42] Elsayed, Medhat, and Melike Erol-Kantarci. "AI-enabled future wireless networks: Challenges, opportunities, and open issues." *IEEE Vehicular Technology Magazine* 14.3 (2019): 70-77.

[43] K. Hamidouche, A. T. Z. Kasgari, W. Saad, M. Bennis and M. Debbah, "Collaborative Artificial Intelligence (AI) for User-Cell Association in Ultra-Dense Cellular Systems," *2018 IEEE International Conference on Communications Workshops (ICC Workshops)*, Kansas City, MO, USA, 2018, pp. 1-6

[44] Y. Shi *et al*., "Machine Learning for Large-Scale Optimization in 6G Wireless Networks," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2088-2132, Fourthquarter 2023

[45] Best Practice Approach for Reducing System Level Energy Consumption in 5G Open RAN". https://symphony.rakuten.com/blog/open-ran-5g-energy-savings-new-best-practices-white-paper-now-available

[46] O-RAN.WG1.OAD, "O-RAN Architecture Description".

[47] M. Merluzzi et al., "The Hexa-X Project Vision on Artificial Intelligence and Machine Learning-Driven Communication and Computation Co-Design for 6G," in IEEE Access, vol. 11, pp. 65620-65648, 2023.

[48] F. Binucci, P. Banelli, P. D. Lorenzo and S. Barbarossa, "Multi-User Goal-Oriented Communications With Energy-Efficient Edge Resource Management," in *IEEE Transactions on Green Communications and Networking*, vol. 7, no. 4, pp. 1709-1724, Dec. 2023.

[49] M. A. Hossain, A. R. Hossain and N. Ansari, "AI in 6G: Energy-Efficient Distributed Machine Learning for Multilayer Heterogeneous Networks," in *IEEE Network*, vol. 36, no. 6, pp. 84-91, November/December 2022

[50] T. Huang, W. Yang, J. Wu, J. Ma, X. Zhang and D. Zhang, "A Survey on Green 6G Network: Architecture and Technologies," in IEEE Access, vol. 7, pp. 175758-175768, 2019.

[51] M.I. Henderson, N. Damir, and L.C. Mariesa, "Electric power grid modernization trends, challenges, and opportunities." *IEEE, Nov*. 2017.
[52] L. Kundu, X. Lin, R Gadiyar, “Towards Energy Efficient RAN: From Industry Standards to Trending Practice”, *arXiv preprint arXiv:2402.11993*. 2024.
[53] R. P.V., "Gauging Carbon Footprint of AI/ML Implementations in Smart Cities: Methods and Challenges," 2022 Seventh International Conference on Fog and Mobile Edge Computing (FMEC), Paris, France, 2022.
[54] van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*, 213 - 218.
[55] United Nations, "Transforming our world: The 2030 agenda for sustainable development," United Nations, Department of Economic and Social Affairs, 2015.
[56] J. Koomey, S. Berard, M. Sanchez, and H. Wong, "Implications of historical trends in the electrical efficiency of computing," IEEE Annals of the History of Computing, vol. 33, no. 3, pp. 46-54, 2011.
[57] Chen, R., Pu, Y., Shi, B. *et al.* An automatic model management system and its implementation for AIOps on microservice platforms. *J Supercomput* **79**, 11410–11426 (2023).
[58] Z. Lyu, H. Wei, X. Bai and C. Lian, "Microservice-Based Architecture for an Energy Management System," in IEEE Systems Journal, vol. 14, no. 4, pp. 5061-5072, Dec. 2020, doi: 10.1109/JSYST.2020.2981095.
[59] MLOps-Standardizing the Machine Learning Workflow. (2021).
[60] M. Liu, S. Peter, A. Krishnamurthy, and P. M. Phothilimthana, "E3: Energy-Efficient Microservices on SmartNIC-Accelerated Servers," in Proceedings of the 2019 USENIX Annual Technical Conference (USENIX ATC 19), Renton, WA, USA, 2019, pp. 363-373.
[61] M. Bagaa, D. L. C. Dutra, T. Taleb and K. Samdanis, "On SDN-Driven Network Optimization and QoS Aware Routing Using Multiple Paths," in IEEE Transactions on Wireless Communications, vol. 19, no. 7, pp. 4700-4714, July 2020.
[62] L. Wang, X. Deng, J. Gui, X. Chen and S. Wan, "Microservice-Oriented Service Placement for Mobile Edge Computing in Sustainable Internet of Vehicles," in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 9, pp. 10012-10026, Sept. 2023
[63] Y. Zhang, Y. Wang and B. Fan, "SDN Based Optimal User Cooperation and Energy Efficient Resource Allocation in Cloud Assisted Heterogeneous Networks," in IEEE Access, vol. 5, pp. 1469-1481, 2017.
[64] N. Chauhan, N. Kaur and B. Fan, "Energy Efficient Resource Allocation in Cloud Data Center: A Comparative Analysis," in 2022 International Conference on Computational Modelling, Simulation and Optimization, December, 2022, doi:10.1109/ICCMSO58359.2022.00049
[65] D. Liao, B. Chen, J. Pan, A. Huang and K. S. Saini, “Resilient scheduling of massive heterogeneous cloud resources considering energy consumption uncertainty” in IEEE International Conference on Frontiers Technology of Information and Computer, November, 2023, doi:10.1109/ICFTIC59930.2023.10455844
[66] M. R. Rezaee, N. A. W. Abdul Hamid, M. Hussin and Z. A. Zukarnain, "Fog Offloading and Task Management in IoT-Fog-Cloud Environment: Review of Algorithms, Networks, and SDN Application," in IEEE Access, 2024.
[67] Chen, L., Chen, Y., Xi, J., & Le, X. (2021). Knowledge from the original network: restore a better pruned network with knowledge distillation. Complex & Intelligent Systems, 8, 709 - 718.

[68] Antonio Polino, Razvan Pascanu, and Dan Alistarh. "Model compression via distillation and quantization." International Conference on Learning Representations. 2018.
[69] S.W. Prakosa, J.-S. Leu and Z.-H. Chen, "Improving the accuracy of pruned network using knowledge distillation", Pattern Analysis and Applications, vol. 24, no. 2, pp. 819-830, 2021.
[70] Y. He and L. Xiao, "Structured Pruning for Deep Convolutional Neural Networks: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 5, pp. 2900-2919, May 2024.
[71] XU Zhiqun, CHEN Duan, HU Zhiyuan, SUN Qunying, Emerging of Telco Cloud, Alcatel-Lucent Shanghai Bell, Co., Ltd., Shanghai 201206, China
[72] Y. E. Gebremariam, D. G. Duguma, H. Y. Park, Y. N. Kim, B. Kim and I. You, "5G and beyond telco cloud: architecture and cybersecurity challenges," 2021 World Automation Congress (WAC), Taipei, Taiwan, 2021, pp. 1-6, doi: 10.23919/WAC50355.2021.9559450.
[73] S. Puhan, D. Panda and B. K. Mishra, "Energy Efficiency for Cloud Computing Applications: A Survey on the Recent Trends and Future Scopes," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), Gunupur, India, 2020, pp. 1-6.
[74] M. Moussaoui, E. Bertin and N. Crespi, "5G shortcomings and Beyond-5G/6G requirements," 2022 1st International Conference on 6G Networking (6GNet), Paris, France, 2022, pp. 1-8
[75] Going Green: Benchmarking the Energy Efficiency of Mobile, GSMA, London, U.K., 2021.
[76] L. M. P. Larsen, H. L. Christiansen, S. Ruepp and M. S. Berger, "Toward Greener 5G and Beyond Radio Access Networks—A Survey," in IEEE Open Journal of the Communications Society, vol. 4, pp. 768-797, 2023
[77] M. Dayarathna, Y. Wen and R. Fan, "Data Center Energy Consumption Modeling: A Survey," in IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 732-794, Firstquarter 2016
[78] M.N. Mahdi, A.R. Ahmad, Q.S. Qassim, H. Natiq, M.A. Subhi, M. Mahmoud, From 5G to 6G Technology: Meets Energy, Internet-of-Things and Machine Learning: A Survey. Appl. Sci. 2021, 11, 8117.
[79] What is the average annual power usage effectiveness (PUE) for your largest data center? https://www.statista.com/statistics/1229367/data-center-average-annual-pue-worldwide/
[80] Google Data Centers, https://www.google.com/about/datacenters/efficiency/
[81] “Commission adopts EU-wide scheme for rating sustainability of data centres” https://energy.ec.europa.eu/news/commission-adopts-eu-wide-scheme-rating-sustainability-data-centres-2024-03-15_en
[82] Shaping Europe’s digital future, the European Commission, 2020. https://commission.europa.eu/system/files/2020-02/communication-shaping-europes-digital-future-feb2020_en_4.pdf
[83] How Germany’s Energy Efficiency Act will impact data center operators, https://www.dentons.com/en/insights/articles/2023/september/25/energy-efficiency-act-relevance-for-data-centers
[84] Hannah Ashai, Data Center Efficiency at the Hyperscale, 2022. http://large.stanford.edu/courses/2022/ph240/ashai2/#:~:text=According%20to

%20data%20from%20the%20Lawrence%20Berkeley%20National,400%2C000%20ft%202%29%20have%20PUEs%20of%201.2%21%20

[85] Q. Li et al., "6G Cloud-Native System: Vision, Challenges, Architecture Framework and Enabling Technologies," in IEEE Access, vol. 10, pp. 96602-96625, 2022.

[86] M. Wiboonrat, "Energy Management in Data Centers from Design to Operations and Maintenance," 2020 International Conference and Utility Exhibition on Energy, Environment and Climate Change (ICUE), Pattaya, Thailand, 2020, pp. 1-7

[87] Xuan-Truong Nguyen, Duy-Anh Dang, Hai-Minh Luong, Three Key-Elements for Data Center Facilities Sizing in Early Stage of Design, 4th Asia Pacific International Conference on Industrial Engineering and Operations Management, Vietnam, 2023.

[88] Z. Cai, "Data Center Technology for Low-Energy Operation and Adjustment of HVAC Systems in Large Gymnasiums," 2022 3rd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2022, pp. 34-37

[89] Cheng Sun, Yun Wang, Michael D. McMurtrey, Nathan D. Jerred, Frank Liou, Ju Li, Additive manufacturing for energy: A review, Applied Energy, Volume 282, Part A, 2021.

[90] The future of greener, more efficient data centers is here, Driving sustainability by reducing data center energy needs with industrial additive manufacturing. Datacenter Dynamics (DCD). https://www.datacenterdynamics.com/en/marketwatch/the-future-of-greener-more-efficient-data-centers-is-here/

[91] Reduce carbon footprint and energy consumption with leak-free liquid cooled heat sinks. Datacenter Dynamics (DCD). https://www.datacenterdynamics.com/en/product-news/reduce-carbon-footprint-and-energy-consumption-with-leak-free-liquid-cooled-heat-sinks/

[92] Modular Data Centers Meet Demands for Speed, Agility and Efficiency, Datacenter Knowledge, 2019. https://www.datacenterknowledge.com/industry-perspectives/modular-data-centers-meet-demands-speed-agility-and-efficiency

[93] Tahseen Khan, Wenhong Tian, Shashikant Ilager, Rajkumar Buyya, Workload forecasting and energy state estimation in cloud data centres: ML-centric approach, Future Generation Computer Systems, Volume 128, 2022, Pages 320-332

[94] Z. Zhang, Y. Zeng, H. Liu, C. Zhao, F. Wang and Y. Chen, "Smart DC: An AI and Digital Twin-based Energy-Saving Solution for Data Centers," NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium, Budapest, Hungary, 2022, pp. 1-6

[95] M. Lin, A. Wierman, L. L. H. Andrew and E. Thereska, "Dynamic Right-Sizing for Power-Proportional Data Centers," in IEEE/ACM Transactions on Networking, vol. 21, no. 5, pp. 1378-1391, Oct. 2013

[96] M. J. Martin et al., "Energy Use in Quantum Data Centers: Scaling the Impact of Computer Architecture, Qubit Performance, Size, and Thermal Parameters," in IEEE Transactions on Sustainable Computing, vol. 7, no. 4, pp. 864-874, 1 Oct.-Dec. 2022