Green Future Networks

Network Energy Efficiency

v1.0

www.ngmn.org

WE MAKE BETTER CONNECTIONS

NETWORK ENERGY EFFICIENCY

by NGMN Alliance

Version:	1.0
Date:	03.10.2021
Document Type:	Final Deliverable (approved)
Confidentiality Class:	P - Public

Project:	Green Future Networks
Editor / Submitter:	**Johan von Perner (Huawei), Vasilis Friderikos (KCL)**
Contributors:	**Javan Erfanian, Bell Canada** **Jianhua Liu, China Mobile** **Saima Ansari, Deutsche Telekom** **Daniel Dianat, Ericsson** **David López-Pérez, Huawei** **Johan von Perner, Huawei** **Vasislis Friderikos, Mischa Dohler, King's College London (KCL)** **Ljupco Jorguseski, TNO NL** **Emre Bilgehan Gedik, Korhan Yaman, Gökhan KALEM, Turkcell** **Ana Galindo Serrano, Maria Oikonomakou, Orange** **Gary Li, William Redmond, Intel** **Marie-Paule Odini, HPE**
Approved by / Date:	**NGMN Board, 4th November 2021**

NGMN e. V.
Großer Hasenpfad 30 • 60598 Frankfurt • Germany
Phone +49 69/9 07 49 98-0 • Fax +49 69/9 07 49 98-41

CONTENTS

1 Executive Summary 5
2 Introduction and Purpose of Document 8
2.1 Introduction 8
3 Definitions 10
4 Energy Efficiency 11
4.1 RAN Site Overview from an Energy Perspective 12
5 How to Decrease Energy Consumption 14
5.1 Base Station Hardware 15
5.1.1 Virtualization of RAN 17
5.2 Processor and Network Server Power Efficiency Improvement 18
5.2.1 CPU Power Management 19
5.2.2 Virtualization Technology 21
5.2.3 Accelerator Use 21
5.2.4 Instruction Set Architecture Improvements 21
5.2.5 Workload Profiling and Optimization 22
5.3 Software & Functions 22
5.3.1 Sleep Mode Functions 22
5.3.2 Symbol Shutdown 28
5.3.3 Channel Shutdown 29
5.3.4 Sparse Antenna Arrays 29
5.3.5 Carrier Shutdown 30
5.3.6 Network Energy Saving Using Artificial Intelligence 30
5.3.7 Network Design 31
5.4 Artificial Intelligence 32
5.4.1 AI in Mobile Telecommunications 33
5.4.2 Energy Savings Through Artificial Intelligence 35
5.4.3 Energy Consumption of AI 36
5.5 Terminal's Impact on Network Efficiency 41
5.6 Sunset of 2G/3G 42
6 Energy Efficiency in Technical Site 44
6.1 Technical Site Cooling 44
6.1.1 Free Cooling 45
6.1.2 Liquid Cooling 45

6.1.3 AI for Technical Site Management 46
6.1.4 Heat Reuse 47
6.2 Next Generation Uninterruptible Power Supply 47
7 Conclusions and Recommendations 49

1 EXECUTIVE SUMMARY

Energy consumption of mobile networks is a key concern for the operators as it not only leads to an increase in the OPEX but has an impact on CO_2 emissions as well. Therefore, Mobile Network Operators (MNOs) are focusing on finding the best possible ways to reduce the energy consumption of their networks either by using latest technology or by optimizing the use of the active and passive components. Introduction of 5G brings more energy consumption due to the deployment of additional radios in new frequency layers but on the other hand the 5G technology is more energy efficient than its predecessors, thanks to the improved spectrum efficiency which comes along with higher Multiple Input Multiple Output (MIMO) schemes and its ultra-lean design. To support the rising use of cellular connectivity in the 5G era, while reducing energy consumption and emissions on a per-bit basis in the context of an absolute reduction in emissions, the mobile networks need to be more efficient. One way to reduce these emissions could be the use of renewable energy sources which is covered in the NGMN "Sustainability Challenges and Initiatives in Mobile Networks" White Paper [1]. To maximize the end-to-end energy performance, it will be crucial for MNOs to adopt a different approach towards network planning, deployment, and management.

Leveraging the spectral efficiency of the 5G air interface and its more advanced sleep modes is important, but further efficiencies can be combined across three levels – the base station equipment level, site level and network level. There is a wide range of techniques that can be used across these three levels of next generation network operation, which are explored in this document. Since the base stations cover the largest part of the energy consumption in a mobile network, this White Paper details various techniques for automatic wake-up/sleep modes including shutdown on symbol, channel or carrier basis and usage of efficient power amplifiers combined with massive MIMO. Since the traffic load varies during the day, it is important to deploy sleep mode functions that shut off hardware when the load is low. Symbol shutdown is the most important function to address this as it saves approximate 10% energy in a less loaded scenario. Complete carrier shutdown can also be used but has less gain and might have impact on user experience. This White Paper explores different sleep mode functions in detail.

Virtualization has been gaining a lot of attention in the mobile industry, which means the decoupling of software from hardware thereby enabling mobile operators to develop and deploy services quickly. This also helps in the agile deployment of the networks and reduces

the dependency on proprietary hardware. Virtual Radio Access Network (RAN) is based on Commercial Off-The-Shelf (COTS) hardware, such as General-Purpose Processors (GPP) and standard Ethernet Network Interface Cards (NICs). With the advancements delivered in modern server platforms, workloads within the wireless and wireline networks can be served with (COTS) servers combined with hardware accelerators for off-loading heavy baseband processing, while at the same time, power saving modes can be achieved without compromising the strict telecom grade determinism requirements.

For the site level, techniques ranging from using renewable energy for on-grid and off-grid sites, smart batteries, power efficient power supplies are explained. Free and liquid cooling are seen as an important solution in technical sites as IT equipment in indoor sites require cooling solutions which currently can represent up to 50% of the network energy consumption. At network level, flexible cooperation between 5G and LTE is necessary to deliver the right amount of capacity at the lowest practical power level in order to plan the network operation based on energy performance.

Massive MIMO is a key technology for improving energy performance as it allows antenna arrays to focus narrow beams towards the users thereby increasing the spectral efficiency. 5G together with massive MIMO is three to five times more spectrum efficient than 4G deployed with traditional radio solution. Furthermore, it is important to select the best antenna configuration depending on the scenario in order to achieve best energy performance. Better scaling of energy consumption can be achieved by switching off parts of the transmitters at lower traffic load.

The device side will be important too since the performance of the device will impact the overall network performance. The device signal strength receiver sensitivity has an impact on the energy performance. With better sensitivity, higher throughput and less re-transmission are needed which means that the base station can use less power.

Many energy saving solutions mentioned in this document have already been implemented in mobile networks. However, these advancements in energy efficiency will not be sufficient as forecasts point to a significant rise in energy consumption over the next couple of years due to considerable increase in traffic across a vast range of use cases, new technologies and spectrum, great deal of connections, and network densification. Here, Artificial Intelligence (AI) could play an important role. By predicting and learning the traffic behaviour, AI algorithms

define the activation/deactivation of sleep mode functionality and site energy management without impacting the overall performance including Quality of Experience (QoE). AI is still in an early phase and more development and research is needed to reach its full potential. AI based energy saving solutions can greatly increase the energy performance of cellular networks.

This White Paper also presents a methodology on calculating the energy consumption of AI. It is based on an approximation which can be refined depending on the specific AI architecture and use case. Also, the analysis is only conducted for a canonical configuration of a CNN, i.e., one convolutional, one pooling and one fully connected layer. The chaining of several convolutional layers scales the energy consumption linearly. Similar energy calculation methodologies are applicable to Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs) and Deep Reinforcement Learning but omitted here for the sake of brevity. A detailed analysis containing the benefits of using AI, its impact on the network, applicable areas, standardization, maturity, and challenges are also listed.

Without these best practices mentioned above, a 5G network despite improved bit/joule power consumption – could typically use more power than a 4G one with similar coverage area, because of the greater density of base stations when higher frequency is used. By adopting the full range of power and site optimisation techniques in current and future networks, mobile operators can reduce or at least keep the energy consumption stable even in a denser network.

2 INTRODUCTION AND PURPOSE OF DOCUMENT

2.1 Introduction

In NGMN 5G White Paper 1 [2], published at the beginning of 2015, we set a goal of an improvement in energy efficiency by a factor of 2000 within 10 years, such that a 1000 times projected increase in traffic can be carried, using half the amount of energy consumption required at the time. Others set a target of at least 100 times improvement of capacity, compared to 4G. The 5G system is significantly more energy efficient than the previous generations, though still in an early phase on its path to achieve the required targets. Furthermore, the end-to-end increase in energy efficiency should focus on sustainability and environmental targets, towards carbon neutrality.

Many trends are contributing to the rising use of data transmitted over fixed and mobile networks. Changes in the way people work, play and communicate have been enabled by modern technology, and continue to drive its usage. Cellular data traffic is projected to grow by 6 times between 2018 and 2024 in emerging economies, and by 3 times in developed markets over the same period [3]. By 2025 there are forecast to be 100 billion connections, including 40 billion smart devices.

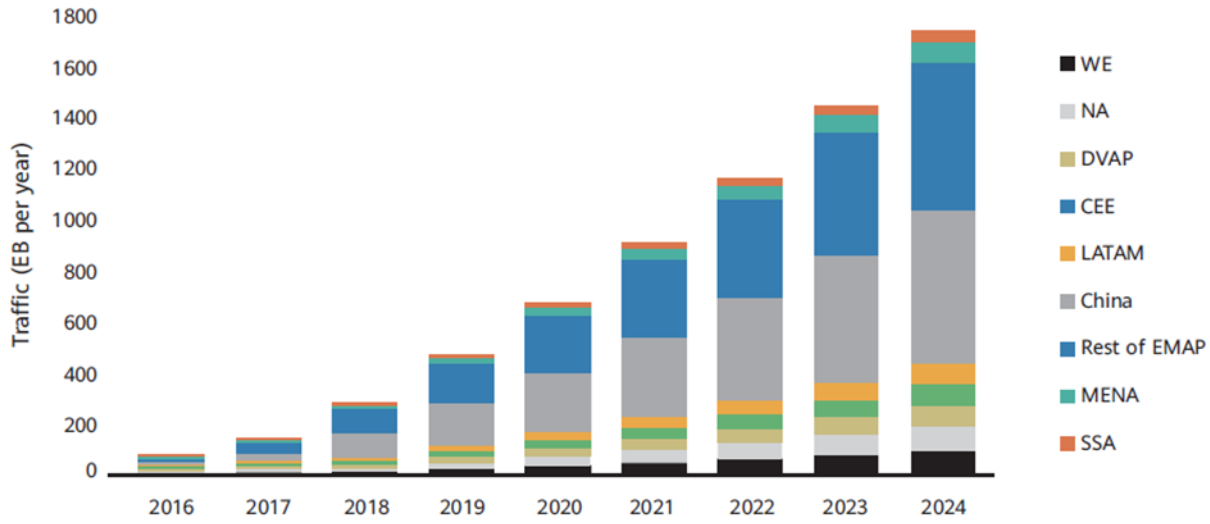


Figure 1: Traffic Growth, Analysis Mason 2020

In order to meet these aggressive targets, a step change in energy efficiency is needed. 5G will play a critical role in addressing this challenge. The most challenging elements of a mobile network, in energy consumption terms, are the radio base stations which represent about 57 percent of the consumed energy [3]. Long Term Evolution (LTE) networks that are highly loaded have a 20 to 30 percent average use of the air interface. Furthermore, approximately 20 percent of the base stations carry 80 percent of the traffic, therefore there are several opportunities to improve the energy efficiency of the base stations. This can be done by reducing the power consumption when no data is needed to be transmitted and increase hardware efficiency, especially when transmission power is below maximum.

3 DEFINITIONS

Throughout this White Paper, the following power and energy related definitions are used:

Power consumption: Amount of energy that is transferred or converted per unit time is called power consumption and it is expressed in SI unit as Watt [W]. Power is calculated by multiplying voltage [V] and current [I]. Other units than Watt are Joules/s or Volt-Ampere [V.A] (1 V.A = 1 W).

Energy Consumption: Amount of power used over a time period is called energy consumption and it is expressed in SI unit as Watt Second [Ws] or Joules [J]. Normally in electricity bills kilowatt hour [kWh] is used (1 Ws = 1 J, 1 Wh = 3600 Ws = 3600 J).

Power Efficiency: The produced power of a unit vs the consumed power for producing the output power per unit time is called power efficiency and normally it is presented as percentage [%]. Power Efficiency = (Output Power / Input Power) x 100 [%].

Energy Performance: The ratio between the produced task or work and the consumed power for producing this task or work over a time period is called energy performance. The task or work could be anything and in telecommunication it can for example be the delivered bits to a User Equipment (UE). In this case the unit could be for example [Mbits / kWh] or [bits / Wh] or [Mbits / Joules]. Since the electricity bills for operators are normally presented in KWh and the work can be expressed as delivering Mbits to a user it would be more convenient to express the unit as [Mbits / KWh].

4 ENERGY EFFICIENCY

The energy efficiency in 5G systems can be attributed to multiple advancements. The enhancements in data transmission, efficiency in control messaging and signalling, and ability to use sleep mode based on traffic and load conditions, are among the capabilities by design. Furthermore, the granular architecture, and increasing disaggregation and cloud native architecture and operation, with virtualization and softwarization, increase agility and reduce footprint. This path towards intelligent and dynamic orchestration, programmability, lifecycle management, and full automation, has great potential to be leveraged towards our goals. In addition, products, and deployment and operational strategies, have already focussed on innovative ways to advance energy efficiency.

Despite these advancements in energy efficiency, the forecasts point to a significant increase in energy consumption, and thereby in CO_2 emissions, in the absence of active intervention, over the next several years. This is due to a considerable increase in traffic across a vast range of use cases, new technologies and spectrum, great deal of connections, and densification. Some of these factors need to be evaluated and optimized, and ultimately lead to the need for intelligent Machine Learning (ML)-based operation. For example, small cells, by nature, introduce efficiency through carrying traffic at lower energy consumption. However, this can have a reverse effect with increasing densification and interference, without intelligent dynamic planning and allocation. Similarly, massive Multiple Input Multiple Output (MIMO) has the potential to increase efficiency, in terms of traffic per unit of energy consumption, if balanced against the complexity and consumption it introduces. Trade-offs may need to be considered to optimize and maintain design goals.

Advancement in equipment power efficiency and network energy performance must be considered from an end-to-end perspective. These can broadly include:

- Specifications and design, such as those related to data transmission, signalling, sleep modes, distributed architecture, cloud, and re-configurable Radio Access Network (RAN) and Edge.
- AI-driven cognitive and autonomous architecture and operation, that support energy efficient dynamic planning, deployment, resource allocation, monitoring and optimization, shutdowns, etc.
- Power efficiency in equipment, devices, boards, and site or data centre cooling, etc.

- Energy harvesting and transfer, and efficiency achieved from circular economy.
- Indirect resources such as external logistics, and end of equipment life.
- Ecosystem, regulatory, and supply-chain collaborative focus, awareness, and organizational alignment.

Mobile Network Operators (MNOs), their partners, and the entire ecosystem will not achieve the sustainability targets, without an end-to-end collaborative, committed and orchestrated effort. At its highest level, the goal involves the inter-connected pillars of achieving digital transformation with full 5G and subsequently 6G realization, sustainability, and social responsibility. With the wide and growing range of use cases, particularly for automated industries, a great impact on other sectors is well expected and articulated.

4.1 RAN Site Overview from an Energy Perspective

A radio access network (RAN) consists of a large number of sites, which typically accommodate two types of equipment: site infrastructure equipment (site) and main equipment (base station). Site infrastructure typically consists of rectifiers for converting AC power to DC power, power backup equipment (e.g., batteries) and other equipment such as air conditioners for cooling. The main equipment is the base station equipment in the cabinet, typically baseband unit and radio units. From the base station, data is transmitted over the air interface (via radio) to the user terminal.

Figure 2 provides an illustration of this and the typical equipment involved, and it also illustrates the energy flow from the main AC input from the grid, via DC power conversion, delivery to the main equipment (base station), conversion to cabinet-top power by the base station, and finally transmission over the air interface to the user terminal. As shown in the figure, the energy flow can be divided into three stages:

Site infrastructure: from the AC main supply to the DC power supply for the base station. The energy efficiency of the site infrastructure, Site EE, can be measured by dividing the DC input power of the base station (P_{BS}) by the AC input power of the site (P_{AC}). This measure is the inverse of the well-known P_{UE} (Power Usage Effectiveness), which is commonly used for data centres.

Base station: from the DC power input (P_{BS}) to the cabinet-top power output of the base station antenna (P_{output}). The power efficiency of a base station can be measured by dividing the cabinet-top power P_{output} by the DC input power P_{BS} of the base station.

Air interface: indicates the link from the output of the antenna on the top of the cabinet to the radio transmission received by the user terminal over the air interface. The energy performance of the air interface, Radio EP, can be measured by dividing the service provided by the base station (e.g., delivered bits, coverage, or number of subscribers served by the base station), S_{pi}, by the output power at the top of the cabinet (P_{output}).

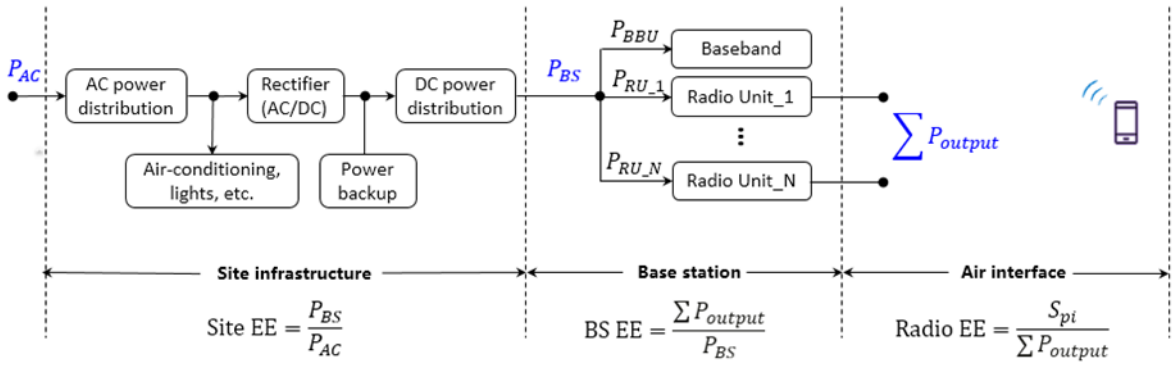


Figure 2: Illustration of a typical RAN site, and energy flow from main AC input to reception at the user terminal.

The overall energy efficiency is affected by these three factors: power efficiency of the site infrastructure, power efficiency of the base station equipment, and energy performance of the air interface. By multiplying these three factors: the Site Power Efficiency (Site PE), the Base Station (BS PE), and the Radio Energy Performance (Radio EP), we obtain the overall energy performance (EP) as:

$$\mathrm{EP} = \frac{P_{BS}}{P_{AC}} \times \frac{\sum P_{output}}{P_{BS}} \times \frac{S_{pi}}{\sum P_{output}} = \frac{S_{pi}}{P_{AC}}$$

When integrated over a time period, we get the Energy Performance as defined in chapter 3, measured in, e.g., bits/Joule or more common used Mbits/Wh.

5 HOW TO DECREASE ENERGY CONSUMPTION

The base stations in a mobile network represent the largest part of the energy consumption, about 57% of total power usage of a typical cellular network is reported by Analysis Mason [3]. By 2025, that figure can be lower when 5G becomes more widely deployed, but still the base stations will still be the biggest consumer of energy. For a converged operator, the RAN could account for around 50% of its total network energy consumption across fixed and mobile networks in 2025, according to a study by three European universities [4], see Figure 3.

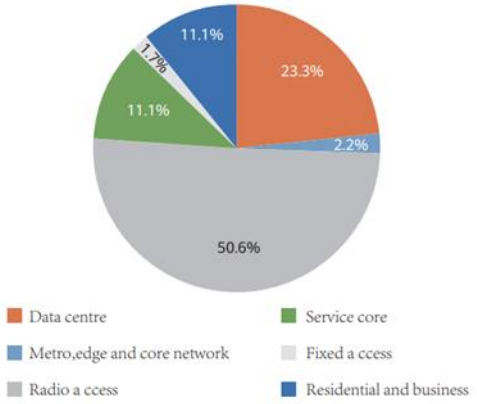


Figure 3: Energy consumption breakdown by network element 2025, University of Split.

5G can be used to improve energy efficiency. More efficient power amplifiers have been developed, renewable energy sources for powering on-grid and off-grid sites, including solar power, are starting to be widely adopted. Moreover, new generation of batteries are becoming an integral part of any 5G site to enhance energy management, and liquid cooling is being implemented to reduce the need for air conditioning, see chapter 6.

3GPP NR specification has enabled a number of new technologies that can help to improve energy performance on network level. The unique 3rd Generation Partnership Project (3GPP) New Radio (NR) power saving technologies are:

- Massive multiple-input multiple-output (massive MIMO)
- Lean carrier design
- Improved Sleep Modes
- Artificial intelligence (AI)

5.1 Base Station Hardware

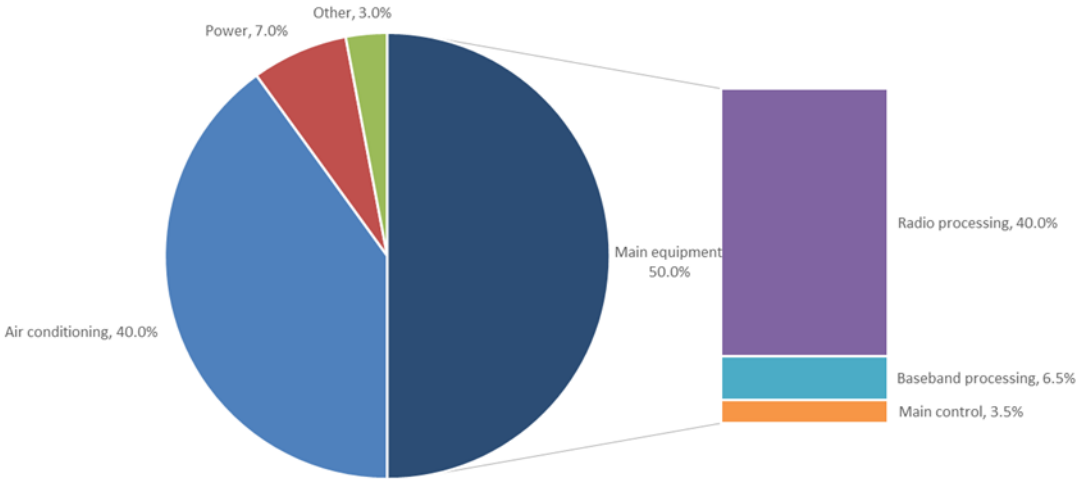


Figure 4: Power consumption of base station.

The largest power consumer of a site is the main equipment consisting of the radio unit, baseband and main control, which accounts for approximately 50%. The second is the air conditioning, accounting for 40%. When looking at the base station, the radio processing accounts for 40%. This unit converts the digital signal from the baseband into amplified radio waves. Since the power amplifier of the radio unit consumes most of the power, it defines the efficiency of a radio unit.

These energy consumption percentages may vary depending on the Telecom equipment power efficiency, the technology and capacity of air conditioning units, the climate and the location of the base station etc. Operators should evaluate their energy consumption percentages by measuring energy consumption for their network before deciding where to focus on for energy savings.

The power consumption of 5G base station using massive MIMO can be divided into two major parts: the antenna unit and the baseband unit (BBU). The power consumption of antenna unit accounts for about 90% of the total consumption of the base station, and it is the main component of the power consumption of the base station. The antenna unit power consumption can be divided into power amplifier, small signal, digital intermediate frequency and power supply.

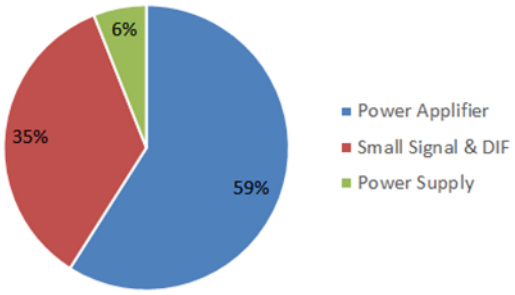


Figure 5: Power consumption of antenna unit.

When 2G was introduced, the power efficiency of the power amplifier was below 20%. Over time efficiency of power amplifiers has been gradually improved due to the improvement of power amplifier architecture design and semiconductor material technology. More advanced Doherty architectures combined with envelope tracking and Gallium Nitride (GaN) materials have been applied in the power amplifiers, increasing the efficiency of power amplifiers more than 50%. Gallium Nitride power amplifiers are widely used in 5G equipment, so the power efficiency is much higher than 2G/3G/4G equipment running in the current network.

The power consumption of the base station changes with the traffic load, and the power consumption ratio of each functional module also changes accordingly. Under the full load condition, the power consumption of the power amplifier accounts for the highest proportion, about 59%; under no-load conditions, the digital intermediate frequency part has the highest proportion of power consumption, about 46%. Therefore, in the field of equipment-level energy-saving technology, it is not only necessary to improve the efficiency of the power amplifier, but also to reduce the basic power consumption of small signal and digital intermediate frequency modules under the condition of low initial load of 5G.

For a traditional radio unit, the power amplifier accounts for the largest part of the power consumption. However, with the deployment of massive MIMO (e.g. 64T64R), power consumption for the digital components will increase due to increased number of transmitters, since more base band processing, e-CPRI processing and support of algorithms like Digital Pre-Distortion (DPD) and Crest Factor Reduction (CFR) is required. Therefore, it is important to use power efficient hardware such as Application Specific Standard Products (ASSPs) and Application-Specific Integrated Circuits (ASICs) and to make sure that the silicon

design or next generation programmable radios can power off modules at low or medium load.

5.1.1 Virtualization of RAN

Decoupling of software from the hardware in the ICT industry and more specifically the mobile core network has been going on for several years. Data centres are becoming more important as cloud services are growing. It is anticipated that this trend will continue, as 5G will address more vertical industries and enterprise businesses.

Virtualization of 5G RANs has been gaining a lot of attention lately which means the implementation of RAN is based on Commercial Off-The-Shelf (COTS) hardware, such as General-Purpose Processors (GPP) and standard Ethernet Network Interface Cards (NICs). Organisations like the O-RAN Alliance and Telecom Infra Project (TIP) are currently developing and promoting for Open RAN. Virtualized RAN is based on disaggregation of hardware and software and can be based on open interface specifications provided by O-RAN or proprietary interfaces. Most common vRAN/O-RAN interfaces are Centralized Unit (CU) - Distributed Unit (DU) split option (F1) and fronthaul split option 7-2x (eCPRI). The baseband unit can be based on GPP COTS leveraging power optimisation techniques discussed in chapter 5.2 and Field Programmable Gate Arrays (FPGAs).

The legacy RAN architecture is based on proprietary hardware and software. Hardware accelerators based on ASICs are used to address power consumption and to provide real time performance. 5G, massive MIMO and advanced interference mitigation function require even more advanced algorithms in order to improve spectrum efficiency.

Chip vendors supporting the development of Open RAN are promoting a baseband solution using GPP combined with NICs for hardware accelerators and may use dedicated accelerators for channel coding and decoding etc. Since massive MIMO systems may have a lot of antenna and baseband processing for layer 1, chip suppliers are now seeking solutions based on low power consumption ASIC or ASSPs for massive MIMO specific tasks, rather than medium or high power FPGAs and GPUs. The power efficiency of ASICs is better than FPGAs and GPPs.

5.2 Processor and Network Server Power Efficiency Improvement

While most of the energy use in networks is in the radio access network, operators also deploy a large amount of general-purpose compute capacity in their core networks, OSS/BSS support systems, and service Data Centres.

With the advancements delivered in modern server platforms, most workloads within the wireless and wireline networks can be serviced with commercial off-the-shelf (COTS) servers, while at the same time, energy efficiency and power saving modes can be achieved without compromising the strict telecom grade determinism requirements. The goals of an energy efficient network can be considered as energy efficiency, efficient data transmission (bits/Hz) and prioritisation of power to key elements to fulfil SLAs. By reconfiguring peripherals such as memory, storage, network and accelerator add-in cards, a COTS system can satisfy requirements for everything from baseband processing in Wireless Access, to User Plane Forwarding in Data Plane Processing, as well as in Control Plane Processing. Common underlying technology allows for flexible deployment using open-source orchestration and management, enabling an energy efficient network.

Most COTS servers are designed to meet certifications such as the US Environmental Protection Agency and US Department of Energy's Energy Star programme, which have been very active in terms of creating specifications for energy efficiency, as well as in promoting energy savings in data centres and server rooms [5]. Similar mechanisms, such as Lot9 regulation and EPEAT registration, exist in Europe, Asia, and the rest of the world.

For example, latest generation COTS systems have demonstrated a performance gain of 1.42X, translating to a 15% energy efficiency improvement for 5G UPF and can achieve twice the Massive MIMO throughput in similar power envelope for vRAN versus prior generation servers [6]. Similarly, a typical COTS servers vendor achieved in 2020 3.2x energy performance on the portfolio since 2015 [7]. These servers are now commonly deployed in 5G cloud native vCore and vRAN to improve 5G network energy efficiency.

While energy efficiency is increasing with each new generation of processors, there is a greater opportunity to implement power saving features and become more energy efficient.

COTS systems provide a range of management functions and APIs to enable data centre or site management software to monitor and react to the utilization of each individual server in a deployment, allowing power-aware scheduling and optimization of power consumption across the deployment [8], [9], [10]. Automation forms a key part of achieving an energy efficient network. Consolidating workloads onto fewer servers, optimizing the cooling approach for each deployment location, and selecting the optimum placement in the network for a workload enable significant reductions in overall energy consumption. Network workloads require fine-grained and fast response to traffic variation. Advanced telemetry combined with the CPU Power Management features described below allows power efficiency techniques to be extended across the network (see [11]).

5.2.1 CPU Power Management

Modern processors, which can be used across multiple segments of wireless networks, have various features and capabilities that can be used to improve energy efficiency. Processors have been increasing in core count and they provide more compute power than previous generations. Given that not all resources may be used all the time, it is important to save power when processing capability is not needed. In this section, we describe technologies that can be leveraged to save power and improve energy efficiency.

5.2.1.1 C States

C-states are idle or sleep states that a multi-core CPU can use to reduce power consumption on a per-core or CPU package level, by powering down portions of the core, package, or both. Disabling portions of the core allows for power savings when the core is not executing instructions. As defined in the Advanced Configuration and Power Interface (ACPI) [28] there are several C states that can be used. C0 is active state, and there are several lower power Cx states. Figure 6 provides an overview of core C states and their power consumption. As the processor core enters deeper C states the power consumed is reduced, but at the same time latency to resume active work increases. Intelligent management C states based on network load conditions can result in power reduction, while not compromising constant network connectivity.

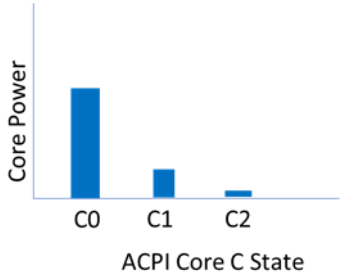


Figure 6: Power Consumption of Core C States.

5.2.1.2 P States

Core Performance States: reducing the core frequency to match the load on the CPU is a common approach when a core is lightly loaded. Modern processors can operate over a range of frequencies (P states), with lower frequencies consuming less power. When traffic is low, selecting a core frequency that does not affect the completion time of workloads can reduce power consumption.

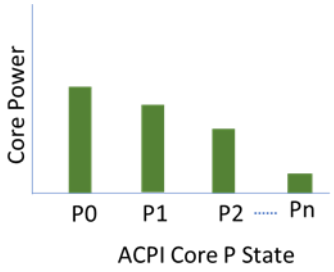


Figure 7: Power Consumption of Core P States.

For example, frequency scaling can be performed as a function of the time-of-day to obtain power reduction. When traffic is high, a processor can run at high frequency and when traffic is light for example during night hours, it can run at low frequency and consume less power but still provide continuous connectivity.

Given high core counts, and the practice of consolidating workloads, processors need to play multiple roles. Frequency tiers can be configured so that frequencies can be matched to workloads for optimum energy efficiency (see [13]).

5.2.2 Virtualization Technology

Server virtualization allows workloads to be optimally scheduled on hardware, for example by consolidating workloads onto a reduced number of CPUs for energy efficiency. With more cores and virtualization technology, multiple applications and workloads can be run on a single server, thus increasing energy efficiency. This allows multiple network workloads to run in virtual machines (VMs) or containers, thus enabling efficient use of the common server resources [14], [15]. Studies have shown that network power consumptions are reduced by 22% and 17% respectively by virtualizing the EPC and RAN (see [16]).

5.2.3 Accelerator Use

General purpose processors are flexible for most workloads. As discussed in Section 5.1.1, custom ASICs can execute fixed workloads in an energy efficient manner, when functionality is well defined and not expected to change over time. For the cases where the assigned function is fixed for a period, but may change over time, FPGAs can be used to execute those functions to get the benefit of higher processing capacity, typically, though, at a higher power cost than ASICs. COTS servers allow integrating ASICs and FPGAs as add-in cards, so that a combination of CPUs and accelerators can be selected to optimize performance and energy efficiency for selected computing workloads. For such add-in cards, the APCI specification defines a set of power states, similar to CPU C states, which can be configured to reduce power usage of the card (see [28]).

5.2.4 Instruction Set Architecture Improvements

When a particular problem or computational task is well understood, improvements are made in general purpose processor to perform that task more efficiently, which can reduce energy consumption. Processor developers continuously study such tasks and create new instructions that applications can use to speed up computation and thus increase performance in given power budget or use lower power states to perform the same work at lower power. Examples of such instructions include Advanced Encryption Standard New Instructions (AES-NI[1]), Advanced Vector Extensions (AVX[2]), and Vector Neural Network

[1] Intel® Advanced Encryption Standard New Instructions (Intel® AES-NI)

[2] Intel® Advanced Vector Extensions (Intel® AVX)

Instructions (VNNI[3]) for artificial intelligence. In networks, the security application VPP IPsec [17] utilises such instructions and is used for secure communication in cloud, edge, and work-from-home use cases. For power savings, instructions such as MWAIT, MONITOR, PAUSE, UMWAIT and TPAUSE are part of the CPU's suite of capabilities which allow the CPU to enter low power C-States [18], [19].

5.2.5 Workload Profiling and Optimization

The COTS ecosystem provides a wide range of tools and services to profile and optimise performance and energy efficiency. Performance [20] (Perf) software allows users to profile a workload so that performance bottlenecks can be identified and removed, resulting in increased performance and improved energy efficiency. Other open-source tools, such as CPU power and Turbostat, allow developers and users to understand P state, C state, and power usage of CPUs.

Once the user understands power consumption and performance by profiling, a combination of these technologies (P-States, C-States, accelerators, virtualisation and power saving instructions) can be leveraged to increase energy efficiency.

5.3 Software & Functions

Typical wireless network energy saving technologies for 5th-Generation networks include silence from symbol level, physical channel level, and machine level. The impact on the SLA is also different, here we provide some analysis and suggestions.

5.3.1 Sleep Mode Functions

Power demands of a radio BS change with the cell traffic load. On one hand, as the cell traffic load increases, the Power Amplifier (PA) progressively becomes the most energy consuming BS component. On the other hand, in no traffic load scenarios, the radio BS power demands are mainly attributed to digital intermediate frequency modules. It is notable that, in no traffic load scenarios and in between control signalling transmissions, BS parts consume energy even though there is no need for transmission. Thus, the opportunity arises to reduce

[3] Intel® Deep Learning Boost (Intel® DL Boost)

unnecessary radio BS energy consumption by progressively deactivating components when they remain unused for transmission.

The idea of shutting down components of a radio BS is already present in 4G. Micro-discontinuous transmission (μDTX) was used in LTE as an energy saving scheme on the radio BS side [21]. According to μDTX, the PA is shut down at symbol level, i.e., for 71 μs, i.e., for at least the duration of a symbol and for multiple repetitions as long as there is no control or data to transmit to the users. As the PA is an energy consuming BS component that has no utility during transmission inactivity, energy saving prospects on the BS side from this feature are important and with no impact on the user side.

It should be noted though that the sound performance of LTE is based on heavy control signalling which is transmitted with a single carrier spacing of 15 kHz and creates radio overhead [22]. This overhead refers to the cell-specific reference signals (CRS) that LTE cells transmit irrespective of their traffic load in order to make themselves detectable to devices and so that link measurements can be executed. These signals are continuously and periodically transmitted over the whole cell coverage, i.e., they are "always on". Importantly "always on" signals such as the Primary Synchronisation Signal (PSS) and the Secondary Synchronisation Signal (SSS), which are used in the initial access procedure and are transmitted once every 5 ms. A device with no a priori knowledge of the cell properties, should search for the PSS/SSS in order to connect to it. Apart from the CRSs, there are also the Channel State Information Reference Signals (CSI-RS) in LTE. CSI-RS are transmitted in a more flexible way than CRSs and they aid channel sounding processes, e.g., interference estimation and multi-point transmission. The overhead created by these signals sets an upper limit in the achievable network energy efficiency, despite the use of energy saving features.

In a 5G system, there are more opportunities to activate energy saving features, such as shutting down of BS components. This happens not only thanks to the lean design of the NR carrier, but also because multiple procedures that create the described heavy signalling overhead of LTE are revisited in NR.

In contrast to LTE, NR supports multiple numerologies as it is designed to support various deployment scenarios with cells of carrier frequency from sub-1GHz up to mm-Wave and with very wide bandwidth allocations [22]. The "always on" signals of LTE, PSS and SSS of

periodicity 5 ms, are now transmitted together with the physical broadcast channel (PBCH) as a synchronisation signal (SS) block with a periodicity varying in the range between 5 ms up to 160 ms. However, the default periodicity is 20 ms for the initial cell search. This in turn means that Stand Alone (SA) cells need to adopt a cell defining SSB periodicity lower or equal to 20 ms in order to be detected during the initial cell search procedure of a UE.

The SS blocks are also used for radio resource management (RRM) measurements in idle, inactive and connected modes of the device [22], [23]. NR can additionally use CSI-RS for RRM measurements in connected mode which differs from LTE where the CSI-RS use is restricted to CSI acquisition. A new SS block measurement timing configuration (SMTC) window has been introduced to notify devices of the SS blocks periodicity and timing that they can use for RRM measurements in idle, inactive and connected modes. The SMTC concept relies on a fully synchronized network for SSB periodicities higher than 5 ms. As a result, the SS blocks of NR are used instead of LTE CRS/PSS/SSS with a longer transmission periodicity for initial access and synchronization. All the other reference signals, e.g., tracking reference signals (TRS), which play the same role as CRS in NR, are configurable in connected mode.

Based on the above, the lean NR carrier allows the activation of energy saving features for multiple granularities of time in between the scarce control signalling and when cell traffic load is zero. In the case of shutting down radio BS components, the longer this cell inactivity is, more radio BS components can be shut down deeply in time, depending on the time needed for their wake up. Thus, energy efficiency prospects at the radio BS increase. The adaptive deactivation of radio BS components for multiple granularities of time will enable sleep mode of different levels.

An indicative starting point of enabled 5G radio energy performance which can be improved with sleep mode is listed in Table 1.

Table 1: Comparison of energy performance for a typical 4G (2T2R, carrier aggregation of 4 bands, 20MHz bandwidth) and 5G (64T64R, 3.5GHz, 100MHz bandwidth) configuration at 30% business load

Technology	4G	5G
Energy Performance	0.06 Mbps/W	0.25 Mbps/W

Table 1 provides a comparison of energy performance for a typical 4G (2T2R, carrier aggregation of 4 bands: 800MHz/1.8GHz/2.1GHz, 2.6GHz, 20MHz bandwidth each) and 5G (64T64R, 3.5GHz, 100MHz bandwidth) radio BS configuration (radio equipment and baseband unit) at 30% traffic load, for a single sector, and taking into account only the actual spectral efficiency. The values are issued by lab measurements.

5.3.1.1 Different sleep mode levels

Thanks to the definition of so-called lean carrier radio access of NR which allows for configurable signalling periodicities, we can better define different levels of sleep modes ranging from deactivation of some components of the base station for several micro-seconds to switching off of almost all of them for one second or more. On a research level, four different levels of sleep modes with a minimum sleep time duration can be extracted [24].

1. 1st level of sleep mode:
   Similar to the μDTX energy saving feature of 4G, this level of sleep mode foresees primarily the deactivation of the radio unit power amplifiers for the duration of at least 71 μs.
2. 2nd level of sleep mode:
   This level of sleep mode foresees the deactivation of more BS components than the 1st one for the duration of at least 1 ms, i.e., one TTI. Components to switch off at this level may be retrieved from the radio frequency (RF) integrated circuit or from the digital processing unit.
3. 3rd level of sleep mode: This level of sleep mode foresees the deactivation of more BS components than the 1st and 2nd ones for the duration of at least 10 ms, i.e., one frame. Similar to the 2nd level of sleep mode, components to switch off at this level may be retrieved from the RF integrated circuit or from the digital processing unit.
4. 4th level of sleep mode:
   This level of sleep mode foresees the deactivation of more BS components than all the previous ones for the duration of approximately 1 s. It is worth noting that, as SS block periodicities vary from 5 ms up to 160 ms, this level of sleep typically cannot be used without loss of connectivity. Deep sleep features like cell switch off, and massive MIMO muting can fall into this category provided that an alternative traffic offload solution, e.g., roaming, exists.

The afore-described sleep mode levels can also be observed in Table 2, along with the indicated minimum sleep times and depths.

Table 2: Sleep mode levels with minimum sleep time and depth required for their activation

Level of sleep	1 (4G like)	2	3	4
Sleep time (Minimum duration)	71 μs	1 ms	10 ms	~1 s
Sleep depth (Components to switch off)	Power amplifier (PA)	PA & more(*)	PA & more(*)	PA & more(*)

An indicative example of how a user coordinates in the time domain with a system where sleep modes are enabled can be observed in Figure 8. According to the figure, after the initial active state, the cell enters a time period when there is neither user activity nor need for signalling transmission. Thus, the BS equipment progressively transits into deeper sleep levels. As soon as there is a user arrival, a transition from the 3rd level of sleep mode to active state can be observed.

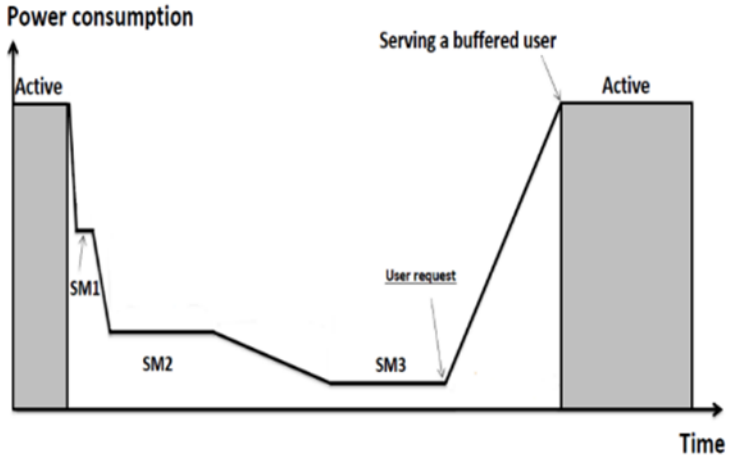


Figure 8: Indicative sleep mode coordination

It is worth noting that multiple long SS block periodicities, which enable deep modes, may have a negative impact on the user experience. Devices performing cell re-selection and handover, might experience delay for RRM measurement, based on the SS block of the NR cell. Moreover, they will detect with a very low probability a cell with SS block periodicity higher than 20 ms since the default periodicity assumed by a device is 20 ms. In an LTE system, this is not an issue as the PSS/SSS are always transmitted periodically every 5 ms.

However, the introduced delay can be calibrated thanks to a sparse synchronisation raster that NR supports [22]. As many frequency bands with wide bandwidth can be deployed in NR, the device would need to spend unreasonable time in a cell search process along all possible carrier positions, as is the case in LTE. The sparse synchronisation raster specifically defines the locations in the frequency domain on which a device must search for the SS block. These locations not always coincide with the centre frequency, as in LTE. However, as soon as the device detects an SS block, it will receive all the system information it needs so as to establish a connection to the cell from the PBCH and the SIB1, including the information about the subcarrier spacing used for the transmission.

Moreover, a 5G NR SA cell with longer than the 20 ms default SS block periodicity could be discovered for cell re-selection in idle/inactive mode, if the right SMTCs are indicated in the serving cell system information (SI). Two SMTCs in idle/inactive mode are needed and would coexist in the network in this case: the one for SA cells with SS block periodicity of 20 ms, which can be detected at initial cell search and the other one for sleeper cells with SS block periodicity higher than 20 ms, which can be detected during cell re-selection. The second SMTC was introduced in Rel. 16 for this functionality [23], [24], [25]. The same principle of two SMTCs was also standardized in connected mode for intra-frequency handover. Other details on the 5G system behaviour that allows discussion on sleep modes can be found in 3GPP specifications [26] and [27].

5.3.1.2 Towards the 5G Network Roll Out with Sleep Mode Functions

Given that sleep modes are linked with both energy savings and some experience of delay on the UE side, a good trade-off between the two aspects is important to be considered for each level, in the roll out process of a 5G radio network:

- With the 1st level of sleep modes, the minimum sleep time for the radio BS is short with no impact on the network performance. The feature is ready for implementation in the roll out process of the 5G network, promising energy savings that start from an average of 10% of the 5G radio equipment unit energy consumption in low load conditions based on measurements. The 1st level of sleep mode is in a continuous

enhancement process of energy savings with more digital or analogue radio BS components being included in the shutdown.

- With the 2nd and 3rd levels of sleep modes, the minimum sleep time for the radio BS is at intermediate level in comparison to the other two. Energy savings are expected to be higher than with the 1st level, but with an impact on network latency that is non-negligible (0-2 ms, [24]) and cannot be handled as straightforwardly as with the deep sleep features. As a consequence, the trade-off between energy savings and experience of delay with the 2nd and 3rd levels remains a field with open issues and plenty of research opportunities. Given the data magnitude in relation to network performance (e.g., traffic load, latency) and energy consumption that has to be considered, AI is a promising tool to address such issues.
- With the 4th level of sleep mode, the radio BS is foreseen to remain inactive for long time granularities and the impact on the network performance has to be considered. The features of cell switch off and massive MIMO muting, which fall into this category, are ready for implementation in the roll out process of the 5G network. However, a guaranteed fall-back capacity on the radio BS coverage area needs to be ensured, i.e., other carriers should be able to carry the traffic of the impacted cell. Entering a sleep mode for a cell should be well-calculated ahead, but exiting the state should be almost instantaneous in order to accommodate emergencies and sudden spikes in traffic.

The enablement of sleep mode to 5G wireless networks promises to further improvement of the 5G energy efficiency indicated in Table 2 especially at low traffic load conditions. More energy saving levers, such as RF design, antenna integration and chipset optimization, should be developed to improve the 5G energy efficiency even further as the traffic still follows an exponential increase and new applications are data consuming.

5.3.2 Symbol Shutdown

For symbol-level silence, a 1st level of sleep mode, the basic principle is that when the 5th Generation gNB detects that some downlink symbols have no data to send, it turns off the PA and other analogue components, thereby reducing the power consumption of the base station, and basically has no effect on the user's latency. In order to achieve silence for more moments, one implementation method is to collect the packet and send it. For delay-sensitive services, it is necessary to comprehensively consider the latency requirements, set a reasonable cycle of collecting packets, and then shut down the PA in the time domain. By adjusting the number of SSB beams when cells have no traffic or light traffic in a specified period of time, the proportions of symbols for which symbol power saving can take effect can

be increased. During automatic beam adjustment, base station needs to adjust the transmit power of common channels to ensure system coverage. When the load is relatively low, up to 30% reduction of power consumption can be achieved through symbol shutdown, in addition, the current 5G base stations support 4G / 5G dual mode. Since 4G requires more control overhead such as CRSs, (see also section 5.2.1), it will reduce the number of symbols that can be turned off, and the corresponding energy saving effect will also be reduced. Coordination of traffic between 4G and 5G using the same PA is important in order to be more efficient, since it is necessary to make sure that data is transmitted simultaneously in time.

5.3.3 Channel Shutdown

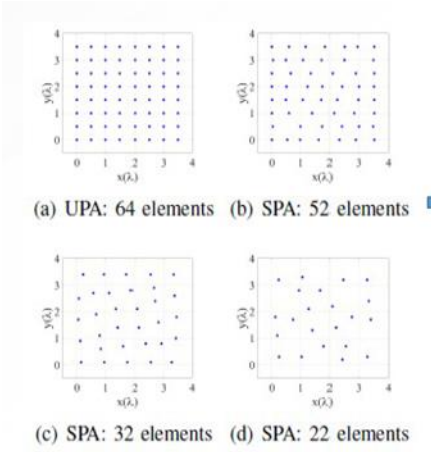
Channel silence refers to the technology of multi-channel base stations such as 64/32 channels by muting some RF channels of the base station with low traffic, thereby reducing the power consumption of the base station. Turning off half of the channels will cause a coverage loss of up to 6dB due to reduced antenna gain and transmitted power, directly causing a significant drop in the throughput of cell edge users. For services with higher throughput requirements, when the user channel environment deteriorates, it is necessary to consider that the coverage cannot be reduced. Power saving between 10% and 20% can be expected.

5.3.4 Sparse Antenna Arrays

Sparse antenna arrays have mainly been used in radar and satellite, but not in mobile communication. This technology has mainly been used to save cost by not using certain antenna elements in a uniform antenna array. With increased number of antenna elements and transmitters, sparse antenna arrays can be an interesting technology for energy saving. By more irregularly silencing certain antenna elements, grating lobes can be smeared out as China Mobile has studied.

Chine Mobile performed system-level simulations in an urban macro environment using an inter-site distance of 350 meters. Data shows the throughput of the unified (64 elements) and the sparse arrays (52/32/22 elements) in Figure 9 below. There exists performance loss in general but rather low when comparing with the reduced energy consumption. When the

number of elements drops to 52 (= 18.75%), the average cell throughput is 212.65 Mbps and the loss is about 11.54%. Meanwhile, there is slight loss in cell-edge user throughput. However, when the number is further reduced to 22 (= 65.63%), both average cell throughput and cell-edge user throughput are challenged by the loss exceeding 20%.



Size of array (elements)	Sparsity factor	Power gained	Average cell throughput	Cell-edge user throughput
64	0%	0%	240.38 Mbps	9.12 Mbps
52	18.75%	18.75%	212.65 Mbps	8.78 Mbps
32	50.00%	50.00%	200.19 Mbps	8.15 Mbps
22	65.63%	65.63%	183.88 Mbps	6.85 Mbps

Figure 9: Example of sparse antenna.

5.3.5 Carrier Shutdown

When the service volume of the entire Base Station (BS) is low during off-peak hours at night the BS energy consumption can be reduced by retaining only the coverage-layer cells and shutting down the capacity-layer cells. The Carrier Shutdown feature periodically checks the service load of multiple carriers within the operator-specified energy saving period. If the service load is lower than a specified threshold, the capacity-layers are dynamically shut down. UEs served by these carriers can camp on or access services from the carrier providing basic coverage. When the load of the carrier providing basic coverage is higher than a specified threshold, the base station dynamically turns on the carriers that have been shut down for service provisioning. When shutting down a carrier, it is important to ensure that full coverage is maintained.

5.3.6 Network Energy Saving Using Artificial Intelligence

The multi-network collaborative energy-saving system using artificial intelligence for energy saving uses basic data such as the configuration and performance statistics of the existing network, and based on the built-in strategy algorithm, activates different types of shutdown

functions under the premise of ensuring the quality of the service to achieve the goal of reducing the energy consumption of the existing network.
Network energy saving using artificial intelligence can be divided into the following phases:

- In the phase of evaluation and design, offline tool platforms are used to quickly analyse energy saving scenarios and pre-evaluate energy saving benefits. In the phase of function verification, the baseline impact of the energy saving solution on Key Performance Indicators (KPIs) and the energy saving benefits are tested.
- In the phase of solution implementation on the entire network, the changes of network KPIs are focused on to ensure that service experience is not affected.
- In the phase of optimization and acceptance, continuous optimization and launch of the online energy saving system maximize network energy saving performance.

AI tools are dependent on the availability of timely and accurate network data in order to assist mobile networks in making correct decisions. Therefore, it is important that activation and deactivation of energy performance features are done without impacting network performance. AI needs to enable decisions based on data collected using different time horizons. The available network KPI's are often limited to longer intervals such as 15 minutes and scheduling of data in the radio is typically carried out using millisecond intervals, therefore the granularity of data used by any AI functionality must be in a way that AI assistance of network making correct decisions does not lead to any network performance degradation. This is very dependent on which level of mobile network architecture the AI tool is designed for.

5.3.7 Network Design

An important strategy for reducing energy consumption at network level is to use the most efficient combinations of spectrum bands available, and to minimise the number of radio technologies in use (through selective activation, or dynamic spectrum usage). The first phase of 5G rollouts is mainly taking place in mid- and low-band spectrum, using a similar site grid to that of 4G. However, as the demand for data consumption continues to rise, some MNOs will move to augment capacity by adding millimetre-wave spectrum in dense hotspots. On a per-bit basis, spectrum in higher bands can be more energy efficient thanks to massive MIMO, but this requires more base stations than when using low bands. Therefore, careful network planning is important to enable MNOs to achieve the optimal balance between capacity and coverage, and to deliver this with the lowest possible power consumption.

With rapid growth of networks, multi-RAT networks are running at the same time and multi-bands are used in many networks. Each band has its own common channels and UE can camp on each band theoretically, which is not energy-efficient. Because of the always-on channels hardware components can't go into deep sleep mode, it is expected that more frequency bands will be employed in future networks. If each band has its specific common channel, the total power consumed can be large even when load is low. To cope with this, a co-design concept is put forward in multi bands in which channels can be distributed in different bands, as shown in Figure 10. One energy-efficient way is that the lowest frequency can be used for UE to camp on. Then SSB and RS channels can be deployed in this band. The higher bands can be only configured with PDSCH and other essential associated control channels. Higher bands can be powered off when there is no traffic. Information of higher bands can be accessed through SIB in the lower band. By doing this, it achieves more power saving as more muting time is realized with deeper sleep mode employed.

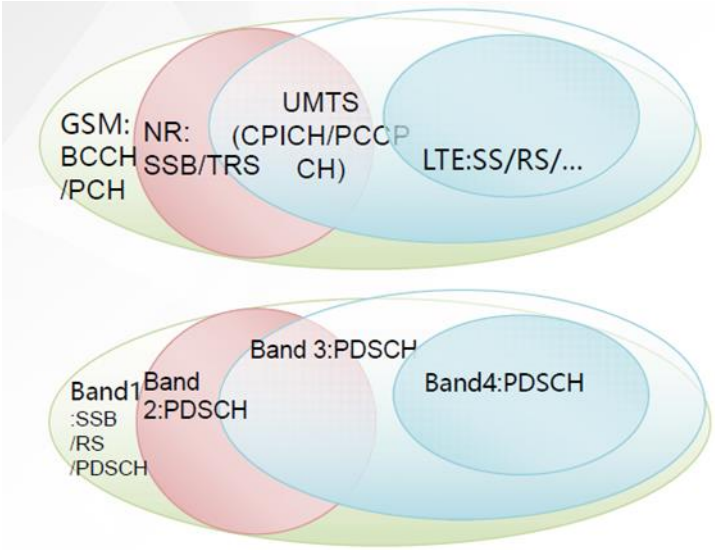


Figure 10: Co-design of Multi-Band.

5.4 Artificial Intelligence

AI has proliferated in recent years, with powerful techniques in signal processing, machine learning and neural networks supporting this quickly growing field. AI is being successfully applied in a variety of technologies, services and applications, including telecommunications.

The aim of this section is to examine the role AI plays with regards to energy performance. The abilities of AI to manage networks such that energy performance is maximized are explored, as well as the energy consumption of AI itself is examined. It is important to understand both so that engineers ensure a positive net balance in terms of energy consumption attributed to AI.

The structure of this section is as follows: the role of AI in telecommunications is elaborated, which standards initiatives have embraced the technology today and a focus on the energy consumption of AI whilst performing a variety of tasks in the network is given.

5.4.1 AI in Mobile Telecommunications

5.4.1.1 Current View of AI in Mobile Telecommunications

5G is a very powerful and versatile technology, encompassing three large use-cases: broadband (eMBB), massive IoT (mMTC) and critical IoT (URLLC). A manual handling at such a high network complexity is almost impossible. Indeed, important and often conflicting KPIs, such as interference, coverage and capacity, need to be optimised prior to the network deployment as well as during operations.

As a result, the mobile telecommunications ecosystem has recognised that automation is needed to ensure the successful scale and operations of telco systems. Indeed, machine-driven decision making has been part of the 3GPP portfolio as of Release 8. For instance, Release 8 introduced eNB self-configuration capabilities with regards to for example automatic physical cell ID assignment; later Releases introduced more automation capabilities, such as automated network optimisation procedures, automated load balancing in heterogeneous networks, among many others.

The algorithmic portfolio is often referred to as Self Organising Networking (SON). In the early years of SON, machine learning (ML) tools, such as Support Vector Machines (SVMs), have been used to optimise network operations. In recent years, however, a new toolset emerged: deep learning based on neural networks. These are able to handle significantly more complex network settings, such as those observed in 5G.

5.4.1.2 AI in Telecoms Standards

As a result, above-discussed automated decision-making capabilities are now embedded in leading industry standards. The most important ones are summarised below:

- **ETSI** Zero-Touch Network and Service Management (ZSM[4]) [: ETSI ZSM was one of the first SDO groups to put forward a standardized view on how AI/ML ought to be embedded into a telco and networking system. The main outputs are ETSI GS ZSM 001 which describes ZSM Requirements; ETSI GS ZSM 002 which defines the ZSM Reference Architecture; and ETSI GS ZSM 007 which provides the glossary of terms and concepts related to ZSM.
- **ITU-T** FG-AI4EE (Focus Group – AI for Environmental Efficiency[5]) [: This Focus Group identifies the standardization needs to develop a sustainable approach to AI and provides guidance to relevant stakeholders on how to operate these technologies in a more environmentally efficient manner to meet the 2030 Agenda for Sustainable Development
- **3GPP[6]**: As already alluded to above, network automation has been of interest to 3GPP ever since R8. However, a general service function was only introduced recently in R16 via the Technical Specification Group Core Network and Terminals in the technical specifications 3GPP TS 29.520 V16.3.0 (2020-03) "5G System; Network Data Analytics Services". A major contribution was the introduction of a standalone core network function named "Network Data Analytics Function (NWDAF)" dedicated to network automation, ML and AI. Similarly, an AI based network function has been introduced in the management plane, "Management Data Function (MDAF)".
- **O-RAN**
  O-RAN has introduced AI/ML in the RIC and defined an API[7] on the A1 interface to manage AI/ML models.

In summary, AI forms an integral part of telecommunications networks today and it is timely to understand its potential with regards to energy savings but also its own energy consumption.

[4] https://www.etsi.org/technologies/zero-touch-network-service-management

[5] www.itu.int/en/ITU-T/focusgroups/ai4ee

[6] https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3355

[7] https://www.o-ran.org/specifications

5.4.2 Energy Savings Through Artificial Intelligence

Power saving techniques discussed in this document need to be implemented across the entire network. This is a huge labour-intensive process. Often, tens of thousands of access points need to be configured, tested, and then monitored. In addition, those spatially distributed access points may be powered by various energy sources. There is a current shift in differentiating the use of the different sources of energy and optimizing consumption based on the relevant type, for example, reducing the amount of energy transferred to a site as well as differentiating the type of energy being consumed. To this end, AI techniques can be used to optimize locally generated renewable energy and energy used from the grid. The study in [28] shows that using ML techniques energy savings of up to 40% and never below 10% can be achieved in a real network (1420 BSs in Milan). Those investigations revealed that ML algorithms that provide both traffic prediction and energy prediction enable utilization of energy, which is produced mostly locally on the site, and only a small amount, between 8% and 23%, is delivered from the power grid. On a larger scale, a joint optimization of the radio resource utilization and the energy utilization for green energy enabled mobile networks is a challenging problem and the use of AI technique can further propel network automation for green operation via energy differentiation.

This is where automation and AI make a significant difference in that it enables the application of energy efficient techniques across networks at scale, in an efficient and effective manner. Another example case study was published by Telenor[8]:

"We had implemented a power-saving feature in our network, which, in short, was supposed to adjust the power usage in all of our 12,000 sectors based on capacity demand. Unfortunately, the feature did not manage to adjust the power saving automatically. To manually set an activation for each of the 12,000 sectors would be impossible considering the time consumption. Setting a fixed power-saving mode for all the sectors would have to be done very conservatively, meaning the loss of huge power saving potential."

Furthermore, it has been demonstrated that an AI-based network automation has shown to entail a significant number of benefits in network efficiency and optimization[9]. Real world testing showed that there has been a 3.6% reduction in power consumption of the radio

[8] Lean, green telco machine: how AI is greening mobile networks - Telenor Group

[9] https://www.gsma.com/futurenetworks/wiki/case-study-ai-use-cases-in-service-assurance

network with AI-based energy saving solution, whilst at the same time there have 17% more user's experienced better streaming through customer-centric coverage and capacity optimization. In overall there has been 2.4% savings achieved across the Radio Access Network (RAN).

AI-ML scenario on 5G networks include:

- Monitoring 5G network power consumption metrics and providing analytics and predictions for service assurance to optimize the placement of VNF (Virtual Network Functions) and containers. This type of analytics can leverage NWDAF with standardized NF load and Slice load analytics & prediction and MDAF in the management plane
- Tracking the UE location and leveraging the NWDAF to analyse and predict UE location and optimising the paging of the device while in sleep mode typically. This is saving energy on the RAN and allows to use device sleep mode more effectively and save battery life on the device at the same time.

5.4.3 Energy Consumption of AI

There is no major prior art on calculating the energy consumption of AI. Below exposure details an approximate working methodology which can be refined depending on the specific AI architecture and use case. Also, the analysis is only conducted for a canonical configuration of a CNN, i.e. one convolutional, one pooling and one fully connected layer. The chaining of several convolutional layers scales the energy consumption linearly. Finally, it is worth pointing out that similar energy calculation methodologies are also applicable to RNNs, GANs, and Deep Reinforcement Learning.

The working methodology focuses on estimating the energy needed during one pass-through cycle in a canonical CNN, i.e., from input to output. This largely captures the energy needed during one cycle of inference, i.e., during the usage of AI. The energy needed for training purposes is not estimated here as we assume training to be done once upfront or rarely, and thus not contributing to the energy expenditure asymptotically.

Once the energy consumption of one cycle is established, various use cases requiring a different number of convolutional layers, input dimensions or cycles can be examined. To this

end, we assume the usage of GPUs which are optimised for operations needed for AI. We commence by calculating the operations per cycle needed:

$$\frac{\text{Operations}}{\text{Cycle}} = \frac{\text{Operations}}{\text{Second}} \div \frac{\text{Cycles}}{\text{Second}}$$

By identifying the operations per cycle of the GPU, this allows us to calculate the number of cycles needed as:

$$\text{No. of Cycles Required} = \text{No. of Operations} \div \frac{\text{Operations}}{\text{Cycle}}$$

To calculate the energy consumption, we not only need the power consumption of the GPU but also the duration of the operation. The duration is obtained as follows:

$$\text{Duration (seconds)} = \frac{\text{No. of Cycles required}}{\text{Frequency (Hz)}}$$

With the power consumption, cycles per second and operations per cycle of a given GPU at hand, we are now able to calculate the consumed energy as:

$$\text{Electrical Energy (Joules)} = \text{Power (Watts)} * \text{Duration (Seconds)}$$

What remains to be obtained is the number of operations a GPU has to perform in a single cycle. The exact analysis is available in [29], and only the results are presented here. Notably, the energy consumption of a single cycle canonical CNN with focus on the convolutional and fully connected layer is presented; the energy consumption of the pooling layer is negligible.

For the sake of exposure, it is assumed here the usage of an Intel UHD Graphics 620 GPU. It consumes 15W and clocks at 300MHz to 1000MHz; it is assumed an average cycle of 650MHz. Furthermore, the GPU is able to handle 403.2 GFLOPS (giga floating point operations per second). This yields approximately 620 GPU operations / GPU clock cycle.

Table 3 and Table 4 show the approximate energy consumption per single CNN passing through of a typical convolutional and fully connected layer in a CNN, conditioned on the

dimension of the spatial input. The input parameters of the second table assume that pooling typically reduces dimensionality by a factor of 100-1000.

Table 3: Energy consumption per single CNN passing through of a typical convolutional layer in a CNN, conditioned on the dimension of the spatial input.

Input	Total number of operations [19]	GPU number of Cycles required	Duration Time (s)	Energy Consumption (J)
100 * 100 input	90,000	$\frac{90{,}000}{620.3} \approx 145.1$ no. of cycles	$\frac{145{,}1}{650\ \mathrm{M}} \approx 0.22\ \mu s$	$15W * 0.22\ \mu s \approx 3.3\ \mu J$
1,000 * 1,000 input	9,000,000	$\frac{9{,}000{,}000}{620.3} \approx 14509.1$ no. of cycles	$\frac{14509.1}{650\ \mathrm{M}} \approx 22\ \mu s$	$15W * 22\ \mu s \approx 33\ \mu J$

Table 4: Energy consumption per single CNN passing through of a typical fully connected layer in a CNN, conditioned on the dimension of the spatial input.

Input	Total number of operations [19]	GPU number of Cycles required	Duration Time (S)	Energy Consumption (J)
100 input neurons, with approx. 20 neurons in the hidden layer	40,100	$\frac{40{,}100}{620.3} \approx 64.65$ no. of cycles	$\frac{64.65}{650\ \mathrm{M}} = 99.4\ \mathrm{ns}$	$15W * 99.4\ \mathrm{ns} = 1.5\ \mu J$
1,000 input neurons, with approx. 50 neurons in the hidden layer	1,001,000	$\frac{1{,}001{,}000}{620.3} \approx 1{,}614$ no. of cycles	$\frac{1{,}614}{650\ \mathrm{M}} = 2.48\ \mu s$	$15W * 2.48\ \mu s = 37.24\ \mu J$

With the canonical energy consumption at hand, the energy consumption of different example configurations and applications can be estimated. For instance, assuming the usage of a popular image recognition CNN, ResNet-50, the amount of convolutional layers is 15. Furthermore, it is assumed that the CNN inference requires a 1,000 x 1,000 spatial input (e.g.,

100 x 100 input layered such that it covers the spatial input of 1,000 x 1,000). A fully connected layer with 100 inputs is assumed. It is also assumed that the inference needs to be performed once a minute in every eNB/gNB.

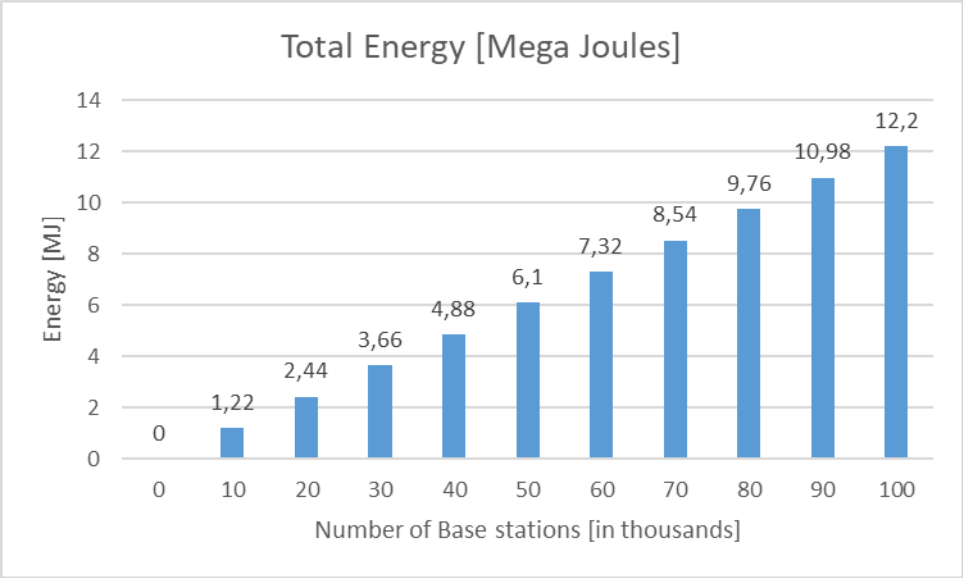


Figure 11: Energy expenditure of AI used at national scale, and assuming an example configuration described in the document.

Approximately, the energy consumption of the AI inference per year per eNB/gNB is 122 Joules (0.03 Wh). The amount of energy consumed as a function of the used eNBs/gNBs is shown in Figure 11. As can be seen, whilst not extraordinary large, the energy consumed by AI is not negligible. Telco networks are expected to use AI in the RAN and other segments in the network. AI is not only used in inference but also may need to be re-trained in regular/frequent intervals. Therefore, the energy consumption shown here is somehow a lower bound and AI systems need to be designed carefully to optimize their energy consumption at each step of the AI model [7]. In other words, the assumption is that the main task is during decision making (inference stage) and it can be assumed that training is only a small fraction of the operational time. In addition, the use of Federated Learning [30], [31], [32] can be explored to further reduce required computations and increase distributed operation to minimize overall energy consumption as well as reduce training/validation requirements.

It is therefore advisable to innovate novel low-energy solutions which will aid networks to become more sustainable.

Table 5: Synopsis of AI techniques in Telecoms

Benefits	• Less (fine) tuning required compared to hand crafted heuristic algorithms. • Data-driven nature renders AI techniques inherently superior in adapting to changing network conditions • Ability to explore & learn complex relationship among different network parameters through training (trial and error)
Impact on the network	• Significant amount of pre-processed data is required for model training, validation and testing of the different deep neural networks • Potentially significant volumes of data transfer between nodes • Added complexity for efficient running deep learning algorithms on dedicated hardware (CAPEX) • Power consumption (OPEX) due to dedicated hardware (GPUs/TPUs/VPUs) for training and for inference
Areas of applicability [10]	• Traffic prediction • Energy source differentiation • Traffic engineering/optimization • Edge caching, content delivery optimization • VNF/SDN network orchestration and workload placement • Mobility management prediction and paging optimization • Advanced BS sleep-modes • Sleep-modes at edge clouds • Network health diagnostics & network monitoring and analysis • Spectrum management (moving beyond cognitive radios) • Radio resource management in NOMA networks • Power control (THz enabled BSs) • Beamforming
Quality of the decision making	• Resource management problems, typically, fall within the regime of NP-hard optimization problems, meaning that optimal decision making cannot be obtained in real (or pseudo real) time • Handcrafted sophisticated algorithms require significant computational time for competitive decision making • Inference time of data-driven AI techniques can be significantly lower compared to sophisticated heuristic algorithms; hence amenable for real-time decision making • Training time data-driven AI techniques requires significant amount of time, however can be done in an off-line manner without affecting operation • Overall performance of AI techniques can surpass achieved decision making of traditional techniques
Maturity	• Current state reflects logical/functional/reference architectures • At trial phase anchoring of AI/ML at different network locations via RAN Intelligent Controllers (RICs) as defined in the O-RAN that provides a centralized abstraction of the network at different time scales (both for real-time and non-real time decision making)
Challenges	• Human in the loop, trustworthiness of decision making • Integration with the network architecture, anchoring of functionalities • Efficient methods to amass data and update • Dataset annotation bottleneck in supervised learning • Applications with small data • Legacy networks • Energy consumption • Operational realities & technical nuances of applying AI techniques on a 'live' network

[10] an indicative rather than an exhaustive list

5.5 Terminal's Impact on Network Efficiency

Normally when talking about a Radio Access Network (RAN), antenna, radio equipment, baseband, site, and other equipment for running the mobile network (core network is necessary as well but not within the scope of a RAN) are considered. It is seldom to take the UEs into consideration as a part of a RAN while we know that without this equipment no mobile network could be run. In general, the RAN's job is to deliver a work/task to what a UE is asking for. These kinds of works/tasks could be for example voice connection, delivering data for different services with different qualities (downloading, streaming and so on). The quality of this work could be very dependent on the channel quality, the available resources, bandwidth, and many other factors. One of these factors is the UE receiver sensitivity which can have a very important roll on how the channel quality by the UE is interpreted and also how a RAN should adapt itself for example in terms of choosing the correct MCS (Modulation Coding Scheme) in order to have an acceptable perception at the UE side.

The UE receiver sensitivity (receiver performance) has a significant impact on RAN coverage and capacity. The better receiver sensitivity the more data throughout across the air interface between a base station and a UE can be obtained which in turn determines the total capacity across the air interface. The receiver sensitivity is dependent on the received SNR, Noise Figure (NF) and the thermal noise (N_o), see Figure 12. The reference sensitivity power is illustrated in Figure 12.

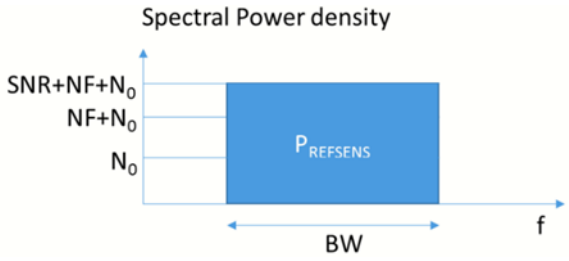


Figure 12: Noise figure.

In cellular system the increase in noise factor at UE causes the reduction of UE throughput or UE coverage due to receiving sensitivity degradation. NF is a measure of degradation of the SNR, caused by components in a signal chain. It is a number by which the performance of the receiver can be specified, with lower values indicating better performance. NF is defined as the ratio of input SNR to the output SNR of the receiver chain. Typically, value of NF for example for an LTE receiver chain is 4 to 5dB.

The receiver sensitivity has a direct impact on which MCS can be chosen by the base station. Higher MCS results in higher payload data and lower MCS results in lower payload data in each packet data over the air interface. It is obvious that more payload data in each packet data over the air interface means a smaller number of packets and hence faster transfer of the total requested data by the UE. On the other hand, less payload data in each packet data over the air interface means a larger number of packets and hence longer transfer time of the total requested data by the UE. Longer transferring time results in more energy being consumed by the base station. By the above description it can be seen that the UE receiver sensitivity has an important impact on the base station energy consumption. Therefore, in a cellular system the increase in noise factor at UE could cause lower energy performance KPI (kbits/kWh) due to receiving sensitivity degradation.

Normally when discussing about which requirement needs to be set to get a higher energy performance RAN the focus will go to only the base station or site and as said not so much or not at all focus on how UEs can impact the overall RAN energy performance. Operators could have a very important role in requesting high requirements on UE receiver sensitivity from UE vendors and make sure that the receiver sensitivities reported by UEs are best in class.

According to a study from 2018 made by Aalborg University [33], the performance of terminals varies a lot. For telephony, the largest variation is for the lower frequency bands with some 12-16 dB variation for the different terminals. For the high bands the variation is some 8-10 dB. The performance variation between left-hand and right-hand usage is also very large for several cases. This shows that the antenna and/or the location of the antenna in some phones is not designed well. For data services, the variation among the different terminals is lower than for telephony, less than 5 dB.

5.6 Sunset of 2G/3G

With the arrival and adoption of 4G and 5G, the question of reducing, or even switching off 2G and 3G networks arose. Currently, 2G and 3G bands are embedded in RAN renewal and therefore benefit from recent more energy efficient equipment. In view of the decrease of the 2G and 3G traffic most operators see an opportunity to reuse 2G and 3G frequencies for 4G and 5G networks, which could mitigate the spectrum scarcity for 5G deployments, avoiding or delaying the addition of new frequency bands. Operators' announcements or strategies

regarding 2G/3G switch-off vary. Some operators already terminated their 2G or 3G services; others announced their switch-off schedules.

In terms of energy efficiency, new equipment deployed could be configured to use less frequency bands for 2G and 3G and more for 4G and 5G. This will imply, in some cases, the need to increase the emitted power. It could be said that the energy efficiency of the network could be improved, i.e., for around the same energy consumption, the capacity of a site is increased.

From an absolute energy consumption point of view, switching off 2G and 3G would allow to decommission only the Base Station Controller (BSC) and Radio Network Controller (RNC) equipment, but might not have a drastic impact on the global site energy consumption.

Concretely operators' positions about this action differ, some of them decide for a stringent decision of completely switching off the 2G and 3G networks. Others prefer to have a more conservative approach and to reduce bandwidth for 2G and 3G to a minimum, i.e., 5MHz for each technology.

Keeping 2G and 3G with a minimum bandwidth for a determined period of time allows to guarantee certain services that still run in those networks:

- Services for the 2G only devices, the 2G/3G devices,
- Voice services for the 4G devices not implementing VoLTE,
- Machine to Machine (M2M) services utilizing 2G,
- eCall services (European Emergency Call system for the cars).

At some point in time either the decrease or the migration of these services will allow to complete the switch off of the 2G/3G networks.

6 ENERGY EFFICIENCY IN TECHNICAL SITE

When evaluating energy efficiency, it is important to consider the whole site installation. Replacing inefficient components with efficient components is key to improve energy efficiency. For example, improve rectifier conversion can increase its efficiency from 90% to 98%. Improved lithium batteries should be deployed as energy storage unit since the maximum operating temperature can be increased from 25℃ to 35℃, reducing A/C power consumption compared with traditional lead acid battery. More modern outdoor cabinet supports innovative cooling technologies such as siphon gravity heat pipe and phase change heat dissipation, improving temperature control efficiency by 80% compared with traditional A/C technology. It also includes having power supply and battery units for outdoor using passive cooling. By moving power supply units closer to the radio unit, cable losses can be reduced. Furthermore, voltage boosting can be used to reduce losses in case of longer cables. A voltage drop can easily contribute to 10% of losses in case of a remote radio solution. More advanced solutions using dynamic voltage boosting to adjust the voltage depending on traffic load can further improve the efficiency of the site. The following sections give a more detailed explanation of different techniques with high energy saving potential for technical sites.

ITU-T published standards covering solution for energy efficiency of sites: L.1210 "Sustainable power-feeding solutions for 5G networks" [34] published also as ETSI ES 203 700, L.1380 "Smart energy solution for telecom sites" [35], L.1381 "Smart energy solutions for data centres" [36] and L.1382 "Smart energy solution for telecommunication rooms" [37]. These documents provide a standardized solution for power feeding, cooling, and monitoring of technical sites.

6.1 Technical Site Cooling

Nowadays, the energy used for cooling can represent as much as 50% of networks energy consumption. Moreover, as power density increases tremendously with the arrival of new generation of servers and Tera routers, the use of air cooling may lead to hot spots in technical rooms and excessive energy consumption.

6.1.1 Free Cooling

By using free cooling units, air conditioner power consumption decreases. In colder regions, by removing air conditioner units, free cooling can be the only cooling equipment in the shelter. Free cooling is one of the alternative cooling methods to reduce energy consumption of air conditioners. Air conditioner units are active cooling equipment that use compressor units to remove heat inside the technical room. Free cooling on the other hand, does not contain a compressor unit. Free cooling is a passive cooling system consisting of a fan, a control unit, a filter, and temperature sensors. The control unit of the free cooling system measures the temperature of the technical room and outside with the help of temperature sensors.

When outdoor temperature is cooler than indoor temperature, the control unit starts the fan unit to take outdoor air into the technical room. In the technical room there is an outlet, close to the ceiling, to remove the hot air inside the technical room. While the fan is working, outdoor air enters the technical room increasing the internal pressure, so excess air inside the technical room leaves through the outlet. When outdoor temperature is warmer than indoor temperature, free cooling does not work. Instead, an air conditioner will start cooling the system room.

A free cooling unit requires less energy compared to an air conditioner unit for cooling the technical room. Also, free cooling fans mostly work on DC energy. In case there is a grid failure or air conditioner breakdown, the free cooling unit works as an emergency cooling system to limit temperature increase in the technical room.

6.1.2 Liquid Cooling

Liquid cooling appears as an interesting alternative, leading to significant energy savings and allowing heat reuse, even in hot countries. Three types of liquid cooling solutions exist, i.e. at cabinet level (to retrofit existing high-density air-cooled equipment), at component levels (with dedicated new equipment) and by immersion in a dielectric fluid.

Liquid cooling at component level is done with heat exchangers called water blocks, installed on the motherboard to cool Central Processing Unit (CPU)/ Graphical Processing Unit (GPU) and memories. Liquid cooling and immersion cooling are very efficient and do not require

any mechanical compressing cooling as liquid at 45°C is sufficient to cool down the electronic parts. It also means that it can work properly even in hot countries.

Liquid cooling solutions aim to significantly lower the network energy consumption by reducing the size and number of cooling machines that use non eco-friendly refrigerant liquids called chillers. Solutions for IT cabinets and water blocks are available and mature for deployment. Some constructors start proposing servers with embedded liquid cooling that can lead to a significant reduction of cooling energy consumption. Some companies are starting to propose immersion solutions. Regarding the radio network equipment, base stations with embedded liquid cooling system, are already available in the market.

Currently available liquid cooling technology in the market is mainly for IT equipment. For optimal performance, IT rooms need to be partitioned to separate equipment cooled with liquid cooling from the legacy equipment.

ETSI publishes standard TS 103 586 covering liquid cooling and ITU-T Study group 5 starts to work on liquid cooling solution for 5G BBU in C-RAN Mode.

6.1.3 AI for Technical Site Management

In data centres many decisions are taken to guarantee good operating conditions. They include control of cooling systems temperature e.g., air conditioning system, free cooling and liquid cooling, and optimization of computing efficiency e.g., computing load repartition, sleeping system, etc. It has been demonstrated that when a smart metering data driven approach is considered, decisions on room temperature computing efficiency have a major impact on the data centre energy consumption.

AI algorithms appear as an interesting solution to administrate these decisions since they can consider complex interactions and follow a global policy on the decision system. They can monitor and analyse the system for a better operational understanding. They can localize breakdowns and might even anticipate some of them. The AI needs to take many decisions to fulfil this objective. Its efficiency will rely on the type and number of actuators as well as on the data associated and how well it represents physical interactions and phenomena.

Concretely speaking, an administrative algorithm in a data centre could take the room temperature and humidity in different spots, air flows inside the technical room, the

dissipated power, and the computational load in servers as inputs. The target would be to control the fan speed, the cold-water production, when to use free cooling and the activation of different sleep modes in severs to minimize the data centre energy consumption. ITU-T L.1305 "Data centre infrastructure management system based on big data and artificial intelligence technology" [38] contains AI applications for cooling and O&M for data centre equipment.

6.1.4 Heat Reuse

Heat reuse is a solution that uses the heat generated by ICT equipment to replace heat needed in other sectors. The basic idea is that ICT generates heat while active and this thermal energy that normally it released into the environment can be reused to heat nearby buildings such as agricultural farms, office districts, etc. Actually, this solution is commonly applied in data centres where the thermal energy can be substantial, and therefore interesting for the deployment of heat reuse systems, especially in countries where environmental temperatures require the use of heating systems for long periods of time throughout the year.

The actual limitations are on one hand the deployment cost of infrastructure to distribute the thermal energy, typically in the form of heat transportation. On the other hand, it is challenging to find a building to reuse the heat near to the source of the thermal energy produced.

ISO/IEC standardize a metric for the data centre to quantify the energy reuse as the ratio between the energy reused outside the data centre and the total energy consumed in the data centre, for details see [39]. Similar metrics could in principle be used for any ICT installation but in this moment the feasibility is limited due to the scale of investment needed for the realization.

6.2 Next Generation Uninterruptible Power Supply

ICT Infrastructure requires an Uninterruptible Power Supply (UPS) to secure the power provision in case of grid breakdown. This back-up power system needs to ensure power supply from several micro-seconds to several hours. Such a dedicated power architecture could be very complex depending on the type and size of the site, for example, in data

centres where especially Alternative Current (AC) is used as power, or in core site, where there is AC and classical 48V Direct Current (DC).

AC architecture introduces losses particularly due to multiple conversions from the UPS input to the electronics inside the IT equipment, and due to losses in the distribution of the building. Total losses could reach 10 to 15%, depending on how old the elements inside the power architecture are.

A new power interface, called 400 VDC or High Voltage Direct Current (HVDC) [40], combines advantages of a DC architecture, by reducing the number of conversion stages compared to AC, with the advantage of "High Voltage" by reducing losses inside the distribution compared to 48 VDC. The gain obtained could reach up to 7%, which is significant considering the high-power consumption of ICT data centres.

In addition to energy savings and implicit carbon emissions and energy bill reductions, the 400VDC reduces UPS CAPEX and fix cost in terms of maintenance, especially compared to AC UPS. It also reduces the volume/surface occupied regarding AC power architecture and the copper used for distribution regarding 48VDC power architecture.

The 400 VDC could also be used in access networks with remote powering solution, to pool battery with the central office, reduce maintenance cost and grid connection cost.

7 CONCLUSIONS AND RECOMMENDATIONS

The energy performance of mobile networks has improved over the years due to introduction of new generations of cellular technology, with better spectral efficiency, advanced hardware with lower power consumption and also many energy saving features introduced in mobile networks. As the deployment of these energy saving features represents an important step in improving energy performance of mobile networks, this White Paper studies the different existing and coming new energy saving features and of course their potential when rolled out in the networks. We have also defined the next steps for further energy performance activities and studies. Other topics addressed by the White Paper are related to the different types of hardware, architectures, and site solutions.

The White Paper also addresses the server virtualization technology which allows workloads to be optimally scheduled on hardware, for example by consolidating workloads onto a reduced number of CPUs for energy efficiency. With virtualization technology, multiple applications and workloads can be run on a single server, thus increasing energy efficiency. This allows multiple network workloads to run in VMs or containers, thus enabling efficient use of the common server resources. Baseband processing of RF signals require a lot of processing power, therefore needs to be implemented in hardware accelerators to be energy efficient.

Data driven AI techniques can not only enable autonomous operation, but the quality of the decision making can help to increase energy performance across different segments of the network. To this end, examples of attainable performance gains in terms of energy consumption have been discussed together with energy consumption of nominal AI techniques that should also be considered.

In view of what has been presented in this paper and to continue improving the energy performance as well as reducing the global networks energy consumption, vendors and operators are encouraged to continue the implementation and activation of advanced sleep mode and shut down features supported by the 5G standard. The path to zero watt at zero load for future network generations is therefore to be continued, especially, considering the use of AI techniques to intelligently coordinate and optimize more precise decisions for activation and deactivation of the sleep-mode and shut-down features, as well as on-demand network dimensioning.

Despite all the work included in the latest communication standards regarding energy saving features, many of these features are not fully implemented or exploited by the operators. This could be explained by the impact on radio KPIs and customer experience that risk a low rating in benchmarking methods. These methods are currently judging operators based on network KPIs such as data rates, latency, and capacity rather than sustainability criteria. To encourage an even larger focus on utilizing periods of low network load to save energy, the industry should work jointly and come up with suggestions on a better rating system for the operators that rewards a balance between traffic performance and energy savings.

Last but not least, the industry is urged to implement optimized cooling and power supply solutions at technical sites in order to minimize the power consumption through techniques such as free cooling, liquid cooling 400 VDC, indoor to outdoor equipment shift, AI for technical site management, etc. Savings performed at this level are a direct benefit for the industry in terms of OPEX and to the environment in terms of energy consumption reduction, and therefore the emissions linked to it.

REFERENCES

[1] NGMN White Paper Sustainability Challenges and Initiatives https://www.ngmn.org/wp-content/uploads/210719_NGMN_GFN_Sustainability-Challenges-and-Initiatives_v1.0.pdf

[2] NGMN 5G White Paper 1 https://ngmn.org/wp-content/uploads/NGMN_5G_White_Paper_V1_0_01.pdf

[3] "Wireless data network traffic forecast, April 2020, https://www.analysysmason.com/Research/Content/Regional-forecasts-/wireless-traffic-forecast-rdnt0/#16%20April%202019

[4] "A survey on recent trends and open issues in energy efficiency of 5G", www.ncbi.nlm.nih.gov/pmc/articles/PMC6679251/

[5] "12 Ways to Save Energy in Data Centres and Server Rooms", [Online]. Available: https://www.energystar.gov/products/low_carbon_it_campaign/12_ways_save_energy_data_center

[6] "Linux kernel profiling with perf," [Online] Available: https://perf.wiki.kernel.org/index.php/Tutorial

[7] HPE Living progress report 2020 [Online] Available: www.hpe.com/us/en/collaterals/collateral.a00113526enw.2020-Living-Progress-Report.html

[8] W. Chen, F. Gao and Y. Lu, "Policy Based Power Management in Cloud Environment with Intel Intelligent Power Node Manager," 2012 IEEE 16th International Enterprise Distributed Object Computing Conference Workshops, Beijing, China, 2012, pp. 66-69, doi: 10.1109/EDOCW.2012.18

[9] "Preserving Performance While Saving Power Using Intel® Intelligent Power Node Manager and Intel® Data Center Manager," [Online]. Available: https://www.intel.pl/content/dam/doc/white-paper/intelligent-power-node-data-center-manager-paper.pdf

[10] "Claim 91,92 5G UPF and FlexRAN," [Online] Available: https://edc.intel.com/content/www/us/en/products/performance/benchmarks/3rd-generation-intel-xeon-scalable-processors/

[11] R. Mijumbi, "On the Energy Efficiency Prospects of Network Function Virtualization," [Online] Available: https://arxiv.org/abs/1512.00215

[12] Greta Vallero, Daniela Renga, Michela Meo, Marco Ajmone Marsan, Greener RAN operation through machine learning, IEEE Transactions on Network and Service Management, vol. 16, no. 3, pp. 896-908, Sept. 2019

[13] R. Bolla, C. Lombardo, R. Bruschi and S. Mangialardi, "DROPv2: energy efficiency through network function virtualization," in IEEE Network, vol. 28, no. 2, pp. 26-32, March-April 2014, doi: 10.1109/MNET.2014.6786610
[14] T. Rokkas, I. Neokosmidis, D. Xydias and E. Zetserov, "TCO savings for data centers using NFV and hardware acceleration," 2017 Internet of Things Business Models, Users, and Networks, Copenhagen, Denmark, 2017, pp. 1-5, doi: 10.1109/CTTE.2017.8260989
[15] "Data Center Manager," [Online] Available: https://www.intel.com/content/www/us/en/software/intel-dcm-product-detail.html
[16] "Speed Select Technology," [Online] Available: https://www.intel.com/content/www/us/en/architecture-and-technology/speed-select-technology-article.html
[17] "Advanced Configuration and Power Interface (ACPI) Specification," [Online]. Available: https://uefi.org/specs/ACPI/6.4/
[18] Veitch. P Browne. J. MacNamara C. Resource Tuning for Energy Efficient Slicing, https://ieeexplore.ieee.org/document/9385531
[19] UMWAIT instruction detail, [Online] Available: https://www.felixcloutier.com/x86/umwait
[20] "Vector Packet Processing IPsec," [Online]. Available: https://wiki.fd.io/view/VPP/IPSec
[21] "Greener, Energy-efficient and sustainable networks: State of the art and new trends, [Online]. Available: www.mdpi.com/1424-8220/19/22/4864/htm
[22] P. Frenger, P. Moberg, J. Malmodin, Y. Jading and I. Godor, "Reducing Energy Consumption in LTE with Cell DTX," 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), Yokohama, 2011, pp. 1-5, doi: 10.1109/VETECS.2011.5956235.
[23] E.Dahlman, S. Parkvall, J. Skold, "5G NR: The next generation wireless access technology", Elsevier Academic press, 2018.
[24] Y. Sano et al., "5G radio performance and Radio resource management specifications," NTT DOCOMO technical journal, vol.20, no.3, p. 79-95, Jan.2019.
[25] F. E. Salem, T. Chahed, E. Altman, A. Gati and Z. Altman, "Optimal Policies of Advanced Sleep Modes for Energy-Efficient 5G networks," 2019 IEEE 18th International Symposium on Network Computing and Applications (NCA), Cambridge, MA, USA, 2019, pp. 1-7, doi:10.1109/NCA.2019.8935062.
[26] 3GPP TS 38.331: "NR; Radio Resource Control (RRC); Protocol specification"
[27] 3GPP TS 38.211: "NR; Physical channels and modulation"
[28] G. Vallero, et al., Greener RAN operation through machine learning, IEEE Transactions on Network and Service Management, vol. 16, no. 3, 2019

[29] D. Mahmood, “Energy Consumption of AI,” BSc Research Thesis, King’s College London, April 2021
[30] Z. Yang et. al., Energy Efficient Federated Learning Over Wireless Communication Networks, IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1935-1949, 2021
[31] “Federated Learning in Mobile Edge Networks: A Comprehensive Survey”; https://arxiv.org/pdf/1909.11875.pdf
[32] Q. Xia, et. al,, A survery of federated learning for edge computing: Research problems and solutions, High Confidence Computing, June ‘21
[33] G. Frølund Pedersen, Aalborg University: “Mobile Phone Antenna Performance 2018”
[34] ITU-T L.1210 “Sustainable power-feeding solutions for 5G networks”, 12-2019, [Online]. Available: www.itu.int/ITU-T/recommendations/rec.aspx?rec=14079
[35] ITU-T L.1380 “Smart energy solution for telecom sites”, 11-2019, https://www.itu.int/rec/T-REC-L.1380-201911-I/en
[36] ITU-L.1381 “Smart energy solutions for data centres”, 06-2020, https://www.itu.int/rec/T-REC-L.1381-202006-I/en
[37] ITU-T L.1382 “Smart energy solution for telecommunication rooms”
[38] ITU-L.1305 “Data centre infrastructure management system based on big data and artificial intelligence technology” 11_2019, [Online]. Available: www.itu.int/ITU-T/recommendations/rec.aspx?rec=14080
[39] ISO/IEC 30134-6 “Information technology — Data centres key performance indicators — Part 6: Energy Reuse Factor (ERF)”.[Online]. Available: www.iso.org/standard/71717.html
[40] ETSI EN 300 132-3 v 2.2.1 (2021-07) Environmental Engineering (EE); Power supply interface at the input of Information and Communication Technology (ICT) equipment; Part 3: Up to 400 V Direct Current (DC), [Online]. Available: www.etsi.org/deliver/etsi_en/300100_300199/30013203/02.02.01_60/en_30013203v020201p.pdf

LIST OF ACRONYMS

3GPP	3rd Generation Partnership Project
AC	Alternative Current
AI	Artificial intelligence
ASIC	Application-Specific Integrated Circuit
ASSP	Application-specific standard product
BS	Base Stations
BSC	Base Station Controller
CFR	Crest Factor Reduction
CNN	Convolutional Neural Networks
COTS	Commercial Off-The-Shelf
CPU	Central Processing Unit
CU	Centralized Unit
DAI	Distributed Artificial Intelligence
DC	Direct Current
DPD	Digital Pre-Distortion
DU	Distributed Unit
GaN	Gallium Nitride
GAN	Generative Adversarial Networks
GPP	General Purpose Processor
GPU	Graphical Processing Unit
HVDC	High Voltage Direct Current
LTE	Long Term Evolution
M2M	Machine to Machine
MIMO	Multiple Input Multiple Output
ML	Machine Learning
MNO	Mobile Network Operators
NIC	Network Interface Cards
NR	New Radio
RAN	Radio Access Network
RIC	RAN Intelligent Controller
RL	Reinforcement Learning
RNC	Radio Network Controller
RNN	Recurrent Neural Networks

TIP	Telecom Infra Project
UE	User Equipment
UPS	Uninterruptible Power Supply
vRAN	virtualized RAN