International Telecommunication Union

ITU-T Technical Report

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(08/2021)

ITU-T Focus Group on Environmental Efficiency for Artificial Intelligence and other Emerging Technologies (FG-AI4EE)

FG-AI4EE D.WG2-02

Computer processing, data management and energy perspective

Working Group 2 – Assessment and Measurement of the Environmental Efficiency of AI and Emerging Technologies

Focus Group Technical Report

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The procedures for establishment of focus groups are defined in Recommendation ITU-T A.7. ITU-T Study Group 5 set up the ITU-T Focus Group Environmental Efficiency for Artificial Intelligence and other Emerging Technologies (FG-AI4EE) at its meeting in May 2019. ITU-T Study Group 5 is the parent group of FG-AI4EE. Deliverables of focus groups can take the form of technical reports, specifications, etc., and aim to provide material for consideration by the parent group in its standardization activities.

Deliverables of focus groups are not ITU-T Recommendations. For more information about FG-AI4EE and its deliverables, please contact Charlyne Restivo (ITU) at tsbfgai4ee@itu.int.

NOTE

This is an informative ITU-T publication. Mandatory provisions, such as those found in ITU-T Recommendations, are outside the scope of this publication. This publication should only be referenced bibliographically in ITU-T Recommendations.

Technical Report FG-AI4EE D.WG2-02

Computer processing, data management and energy perspective

Summary

Technical Report FG-AI4EE D.WG2-02 proposes a set of good practices to improve the energy efficiency of cyber-physical applications – making use of the Internet of things (IoT), artificial intelligence (AI), and digital twins. First, the report introduces the cyber-physical paradigm, engineering reference framework, and a couple of system deployments models. Secondly, it describes three end-to-end use case typologies to be addressed (i.e., monitoring applications using smart IoT systems and AI software; smart applications using edge computing and cloud data centres; and simulation applications using digital twin patterns). Energy efficiency practices are discussed adopting a circular value-chain model that consists of three main steps: data storage; data transfer or movement; and data processing or analytics. Finally, Technical Report FG-AI4EE D.WG2-02 offers a set of recommended practices relating to each component of the three end-to-end use case typologies.

Keywords

cyber-physical systems and applications, energy efficiency, big data, AI, IoT, digital twins.

Change Log

This document contains Version 1 of the ITU-T Technical Report on "*Computer processing, data management and energy perspective*" approved at the ITU-T Study Group 5 meeting held virtually 2021-10-21.

Editors: Stefano NATIVI, European Commission DG JRC, Email: Stefano.NATIVI@ec.europa.eu

Paolo BERTOLDI, European Commission DG JRC, Email: Paolo.BERTOLDI@ec.europa.eu

Tiago SERRENHO, European Commission DG JRC, Email: Tiago.SERRENHO@ec.europa.eu

Table of Contents

Technical Report FG-AI4EE D.WG2-02

Computer processing, data management and energy perspective

1 Scope

This Technical Report presents a set of well-adopted energy efficiency practices for cyber-physical system classes and applications – enabled by artificial intelligence (AI), big data (BD), Internet of things (IoT) and other innovative technologies.

To do so, a set of relevant and significant use cases are first introduced. Second, system classes are identified. Finally, according to a circular value-chain model, system efficiency practices are specified and mapped to components of cyber-physical systems.

2 References

None.

3 Definitions

3.1 Terms defined elsewhere

This Technical Report uses the following terms defined elsewhere:

3.1.1 artificial intelligence [b-ISO/IEC/IEEE 24765]: branch of computer science devoted to developing data processing systems that perform functions normally associated with human intelligence, such as reasoning, learning, and self-improvement.

NOTE – AI systems are designed to operate with varying levels of automation.

3.1.2 big data [b-ITU-T Y.3600]: A paradigm for enabling the collection, storage, management, analysis and visualization, potentially under real-time constraints, of extensive datasets with heterogeneous characteristics.

3.1.3 cloud computing [b-NIST SP 800-145]: A model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

3.1.4 cyber-physical systems [b-NIST FCPS]: Smart systems that include engineered interacting networks of physical and computational components.

3.1.5 data centre [b-ITU-T Y.4051]: Structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of information technology and network telecommunications equipment providing data storage, processing and transport services together with all the facilities and infrastructures for power distribution and environmental control together with the necessary levels of resilience and security required to provide the desired service availability.

NOTE 1 – A structure can consist of multiple buildings and/or spaces with specific functions to support the primary function.

NOTE 2 – The boundaries of the structure or space considered the data centre, which includes the information and communication technology equipment and supporting environmental controls, can be defined within a larger structure or building.

3.1.6 deep learning [b-ISO/IEC TR 29119-11]: Approach to creating rich hierarchical representations through the training of neural networks with one or more hidden layers.

3.1.7 digital twin [b-ISO/TR 24464]: Compound model composed of a physical asset, an avatar and an interface..

3.1.8 edge computing [b-ISO/IEC TR 23188]: Distributed computing in which processing and storage takes place at or near the edge, where the nearness is defined by the system's requirements.

3.1.9 energy efficiency (EE) [b-ITU-T L.1330]: The relation between the useful output and energy/power consumption.

3.1.10 Internet of things; IoT [b-ISO/IEC 20924]: Infrastructure of interconnected entities, people, systems and information resources together with services which processes and reacts to information from the physical world and virtual world.

3.1.11 machine learning [b-ISO/IEC TR 29119-11]: Process using computational techniques to enable systems to learn from data or experience.

3.1.12 sensor [b-ITU-T Y.2221]: An electronic device that senses a physical condition or chemical compound and delivers an electronic signal proportional to the observed characteristic.

3.2 Terms defined in this Technical Report

This Technical Report defines the following terms:

3.2.1 big data analytics platform: Ecosystem of services and technologies that needs to perform analysis on voluminous, heterogeneous and dynamic data.

3.2.2 infrastructure-as-a-service (IaaS): A platform supporting the resources needed by other layers. IaaS can be "programmed" by utilizing provisioning tools. Because of this programming interface, even if IaaS is often (but not only) made of "physical" resources, IaaS can be considered as a component.

NOTE – Based on [b-Sayadi].

3.2.3 platform-as-a-service (PaaS): Systems offering rich environments where to build, deploy, and run applications. PaaS provides infrastructure, storage, database, information, and process as a service, along with well-defined APIs, and services for the management of the running applications, such as dashboards for monitoring and service composition.

NOTE – Based on [b-Veitch].

4 Abbreviations and acronyms

This Technical Report uses the following abbreviations and acronyms:

5G fifth Generation

AI Artificial Intelligence

API Application Programming Interface

BD Big Data

BRR Best Resource Ratio

CPU Central Processing Unit

DSL Digital Subscriber Line

EE Energy Efficiency

GUI Graphical User Interface

IaaS Infrastructure-as-a-Service

ICT Information and Communication Technology

IoT Internet of Things

IT Information Technology

IX Internet exchange

M2M Machine to Machine

ML Machine Learning

NLP Natural Language Processing

ONU Optical Network Unit

PaaS Platform-as-a-Service

PON Passive Optical Network

PoP Point of Presence

PSU Power Supply Unit

PUE Power Usage Effectiveness

5 Conventions

None.

6 Computer processing, data management and energy perspective

6.1 Introduction

By 2023, 5.3 billion people will have access to the Internet, up from 3.9 billion in 2015 [b-Cisco]. Data centres support the information technology (IT) equipment required to provide the services accessed by these billions of Internet users. Ranging from small cabinets to large warehouses hundreds of thousands of square metres in size [b-Shebabi, 2016], data centres are designed to provide reliable access to power, cooling, and Internet connectivity for the IT equipment located within – servers, networking and storage [b-Mytton, 2020a].

By 2021, there will be 7.2 million data centres around the world, down from 8.5 million in 2015 [b-Thibodeau]. This fall is due to the ongoing migration of computing resources to the cloud. In the past, customers bought and owned physical equipment for which they were responsible and deployed space leased from data centre operators. Most growth in usage is now in the cloud, where customers buy units of computation, storage and networking, charged according to usage by the second, hour or per user request. The top three cloud providers by usage – Amazon Web Services, Microsoft Azure, Google Cloud Platform [b-Flexera] – make up the majority of the $236 billion cloud market [b-Adams], and are responsible for some of the largest data centre operations. These hyperscale cloud providers were operating 541 data centres in 2020, with another 176 under construction [b-Synergy].

Estimates of global data centre energy consumption for 2020 range from 196 TWh [b-Masanet] to 287 TWh [b-Hinterman], and there is considerable variance in how this is expected to grow over the coming years. Some projections suggest that global data centre energy has plateaued and will grow by only 5% to 209 TWh in 2023.[b-Masanet] Other projections suggest that data centres in China alone will use 266 TWh of electricity by 2023 [b-Greenpeace] and 96.2 TWh in the EU28 by 2025, 60% from cloud data centres [b-EU cloud market].

Although some of the uncertainty in these figures is due to rapid technological change, such as the introduction of new processors e.g., graphical processing units for machine learning (ML), the growing numbers of IoT devices [b-Shebabi, 2018] and the impact of the end of Moore's law central processing unit (CPU) performance improvements [b-Leiserson], the range in figures also highlights another challenge: transparency. Moving to the cloud means customers no longer have any visibility into the resource consumption of their IT infrastructure [b-Mytton, 2020b]. When customers purchased and ran their own IT equipment, they could directly calculate energy usage and embodied emissions, whereas the data needed to make environmental assessments is not provided by cloud

vendors. This makes it difficult to begin to address data centre energy because the numbers needed to pinpoint areas of focus in the area with most growth are not available.

6.2 Cyber-physical paradigm (IoT, AI analytics, and digital twin innovative technologies)

Cyber-physical systems are smart and include engineered interacting networks of physical and computational components. Cyber-physical frameworks make use of many existing technologies (those for communication networks, information, sensing and control, software, hardware and devices) and combine them to improve operations, lower costs, create new products and business models, and enhance engagement and customer experience. Often, these frameworks are also referred to as IoT or digital twins systems and applications. Figure 1 depicts the archetype of cyber-physical frameworks.

Cyber-physical frameworks enable a very wide spectrum of applications and integrate systems from different vertical sectors (enterprise, consumer, government, industries etc.) [b-Bradford]. Cyber-physical application domains embrace: smart city; smart grid; smart home or building; digital agriculture; smart manufacturing; intelligent transport system; smart energy; digital health; etc.

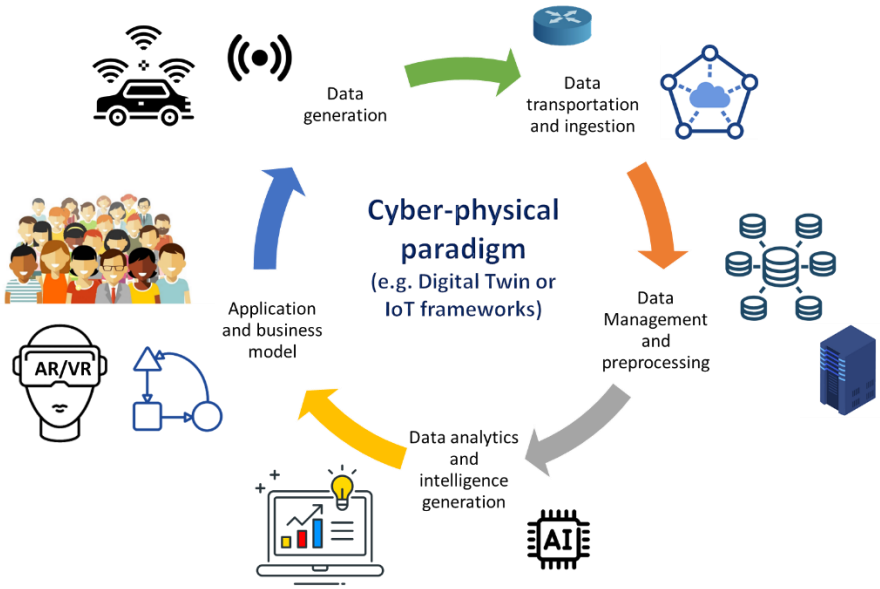

Figure 1 – The archetype of cyber-physical frameworks

6.3 Digital twin pattern

The digital twin communication pattern implements the cyber-physical paradigm; it has been around for several years in the manufacturing sector. Nowadays, due to the digital revolution, it is gaining popularity in other sectors of the economy and society. The pattern is showed in Figure 2.

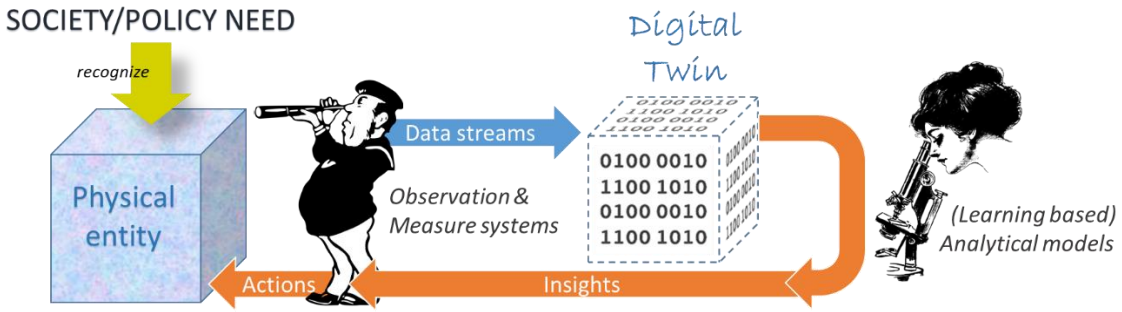

Figure 2 – Digital twin communication pattern

6.4 Cyber-physical reference framework

Mobile technology (e.g., fifth generation (5G) and the forthcoming 6G), cloud computing (e.g., cloud-based data centres and edge computing), BD and deep analytics (e.g., predictive, cognitive, real-time and contextual) play important roles for cyber-physical systems and applications, by gathering and processing data to achieve the final result of controlling physical entities and impacting virtual entities [b-Bradford][b-EC EGAI].

In a general setting, the cyber-physical platform reference framework considers the following main digital components belonging to four different technology tiers [b-EC EGAI] which are depicted in Figure 3:

1 assets/sensors;

2 networks;

3 computing systems;

4 (big) data analytics platforms;

5 software applications.

The archetypical engineering architecture (or reference framework) of a cyber-physical (e.g., digital twin or IoT) platform is shown in Figure 3 [b-ITU-T Y.3502][b-ISO/IEC 30141][b-EC SFC][b-Ferreboeuf][b-EU JRC TR108354].

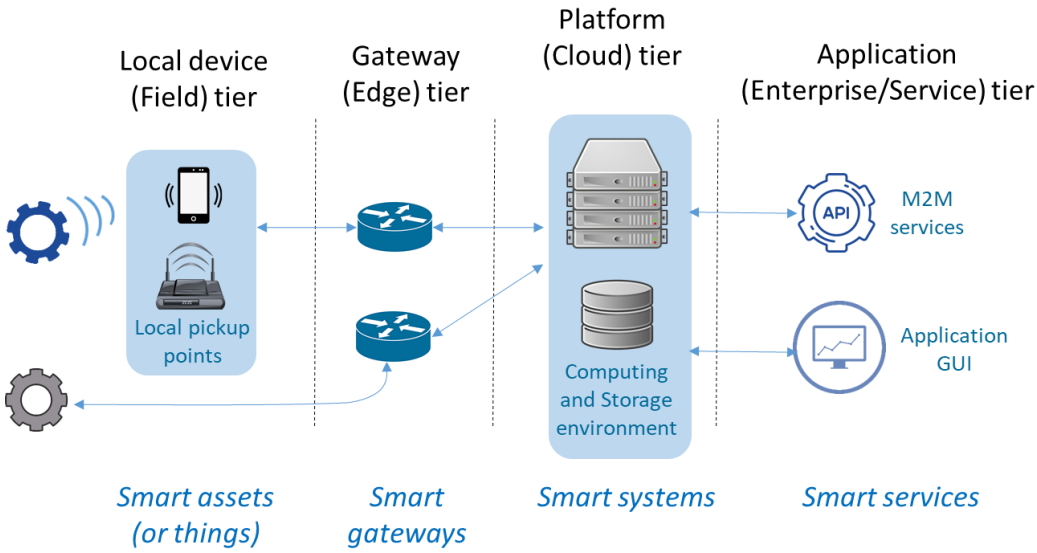

Figure 3 – Cyber-physical architecture reference framework
GUI: graphical user interface; M2M: machine to machine

The field tier consists of smart assets (e.g., sensors), while the gateway (or edge) and platform (or cloud) tiers contain smart gateways and systems, respectively. Finally, the application (or enterprise/service) tier manages smart services [b-EU JRC TR108354][b-Energy Star][b-EC Ecodesign GL].

6.4.1 Smart assets (or things)

The local device (or field) tier comprises different types of devices, ranging from a smart machine to a sensor or an actuator and representing the networked edge node. They are also called smart assets (or things) because they use information and communication technology (ICT) capabilities (such as network, computing and storage) to implement autonomy and collaboration.

6.4.2 Smart gateways

The gateway (or edge) tier contains the edge gateways. They are computing devices (i.e., functional units that can perform substantial computations) that operate as connectors (e.g., by implementing network connection and protocol conversion) between the smart assets (i.e., the physical world local devices) and the digital world. Real-time data analytics are performed by these components. Wherever necessary, for security and transparency reasons, a digital ledger (e.g., by using block-chain technology) may be included.

6.4.3 Smart systems

The platform (or cloud) tier consists of computing servers (and software) that enable non-real-time analytics and manage the cyber-physical system as a whole, by orchestrating the diverse components and the required ICT capabilities, in order to enable the final application or service business logic. These smart systems are commonly constructed based on the collaboration of multiple distributed smart gateways and servers to support elastic expansion of network, computing, and storage resources – see virtual infrastructures and platforms. Cloud-computing and edge-computing servers are typical examples of smart systems.

6.4.3.1 Cloud data centres and edge or fog computing systems

While cloud data centres are large facilities deployed in a limited number of locations (due to special infrastructure and management needs), in a digitally transformed society, cloud users are spread everywhere – IoT and 5G-enabled applications are significant examples. Commonly, clients and users are far from the cloud data centres managed by their preferred providers. Edge or fog computing infrastructure is likely to be closer to those devices and applications to bring computing capacity with lower response time [b-EC Reg 2019/424]. Therefore, in the edge-computing model, an important part of computing and sometimes storage happens at the edge of the network and not in the cloud data centre.

In principle, this allows a reduction in the data quantity to be moved around the network and distribution of the computing load. Edge-computing infrastructures connect the physical and the digital worlds enabling the development of smart systems and applications. While cloud computing effectively supports non-real-time and long-period data driven scenarios, edge computing is effective for real-time and short-period data driven scenarios – such as local decision-making. Commonly, edge computing does not replace cloud services but complements them, reducing storage requirements, decreasing latency, and providing real-time responses to user and application requests.

The edge can be defined in several different ways. Some providers define their edge as a point of presence (PoPs) in major cities. These PoPs may operate a complete copy of the system to offer all functionality, located closer to the user for lower latency and reduced data transfer. These PoPs may be located in major Internet exchanges (IXs), such as LINX (London), AMS-IX (Amsterdam), DE-CIX (Frankfurt), JPNAP (Tokyo). Other providers operate a subset of their platform functionality, such as popular content caching, in a large number of PoPs. These may be deployed in Internet service provider networks much closer to the user, such as the telephone exchange in the nearest town or near to radio mobile telephone towers. For example, Google has three layers to its network and the most granular edge-caching nodes are deployed close to major population centres with multiple nodes within countries, not just in the main IXs [b-Google].

6.4.4 Smart services

The application (or service) tier contains the business logic software that generates and exposes actionable intelligence to cyber-physical system users and clients. The business logic software makes use of smart systems functionalities. Cyber-physical smart services range from observation and monitoring to decision making and simulation.

6.5 Cyber-physical system deployment models

According to the end-to-end application considered, the cyber-physical computing architecture, depicted in Figure 3, can be deployed using either a three- or a four-layer model [b-EC Ecodesign GL].

6.5.1 Cyber-physical architecture three-layers deployment

A three-layer model is common for application scenarios where smart services are distributed, i.e., deployed in one or more scattered areas, each of them characterized by a low traffic volume. Most data processing is done at run-time by the smart gateways and the cloud-enabled service environment is used to enable services distribution and reach user devices. Smart systems are not deployed in a dedicated layer (not much data exchange and secondary processing are needed), but are part of either the gateway or the service layers [b-Bradford][b-EC EGAI][b-EC Ecodesign GL].

Smart assets are processed locally by the smart gateways, which provide real-time streaming data analysis. In addition, the smart gateways aggregate multiple and heterogeneous data streams sending non-real-time data to the cloud for storing and possible additional processing. Finally, each smart gateway implements network services (noticeably, access to and local management of smart assets), security services and small-scale local data storage.

Typical examples of these application scenarios are: smart devices monitoring and control; and smart environmental protection. The three-layer deployment model is shown in Figure 4.

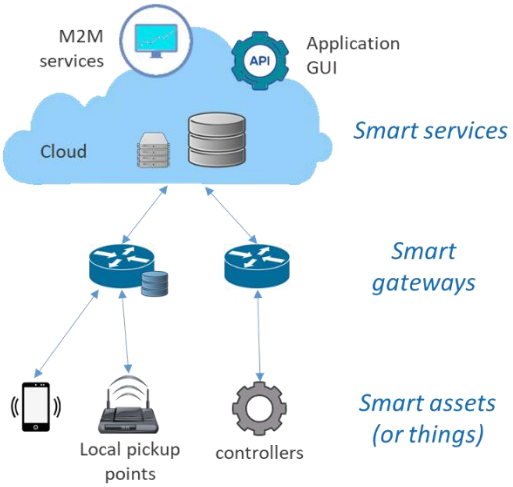

Figure 4 – Three-layer deployment schema of cyber-physical architectures

6.5.2 Cyber-physical architecture four-layer deployment

A four-layer model is common for application scenarios where smart services are deployed centrally, and the traffic volume is high [b-EC Ecodesign GL]. A large amount of data and many local application systems are deployed at the edge of the network. Therefore, it is necessary to provide a large amount of computing and storage resources near the edge – i.e., distributed smart systems. This is achieved by deploying a layer consisting of a set of locally distributed smart systems. They are in charge of aggregating data for secondary processing; primary processing, in real time, has already been done by the smart gateways and the smart assets. The locally distributed smart systems are interconnected to exchange data and knowledge. These (commonly cloud-based) systems support horizontal elastic expansion of computing and storage resources and implement real-time decision-making and optimization operations locally [b-Bradford][b-EC EGAI][b-EC Ecodesign GL]. The service environment (cloud-enabled) is then used to connect with users – see ubiquitous Internet connection.

Typical examples of these application scenarios are: video analysis; distributed grid; and smart manufacturing. The four-layer deployment model is shown in Figure 5.

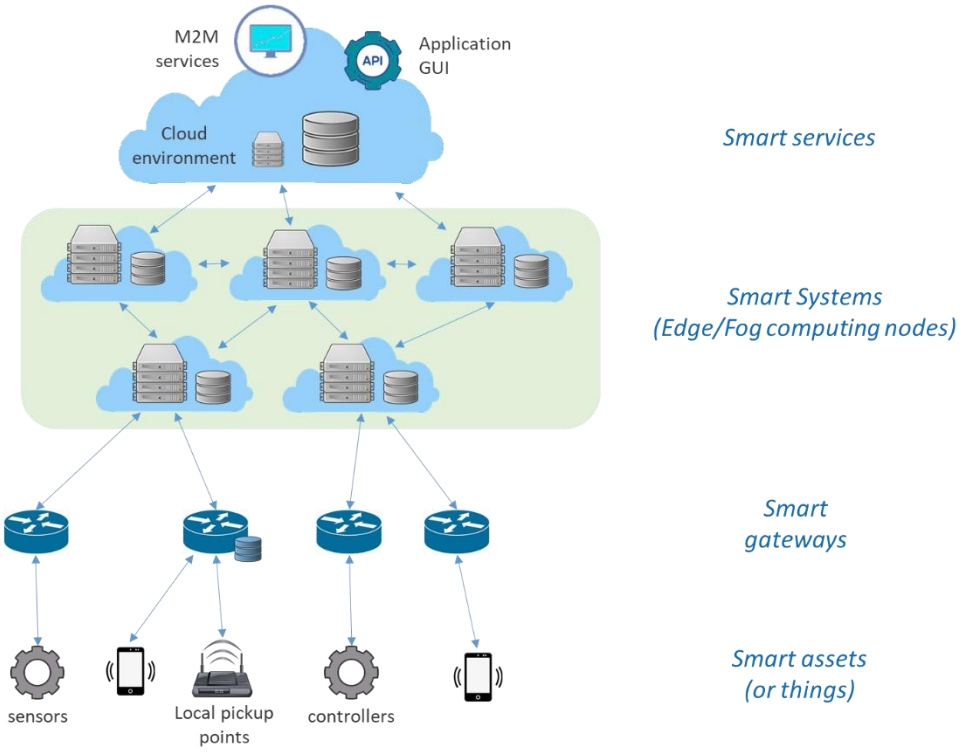

Figure 5 – Four-layer deployment schema of cyber-physical architectures

6.5.2.1 Cloud environment deployment

Cloud platforms can be public, private and hybrid. In particular, a cloud can be seen as an extension of an enterprise data centre (i.e., a private facility operated for the sole use of supporting a single organization) – see Figure 6.

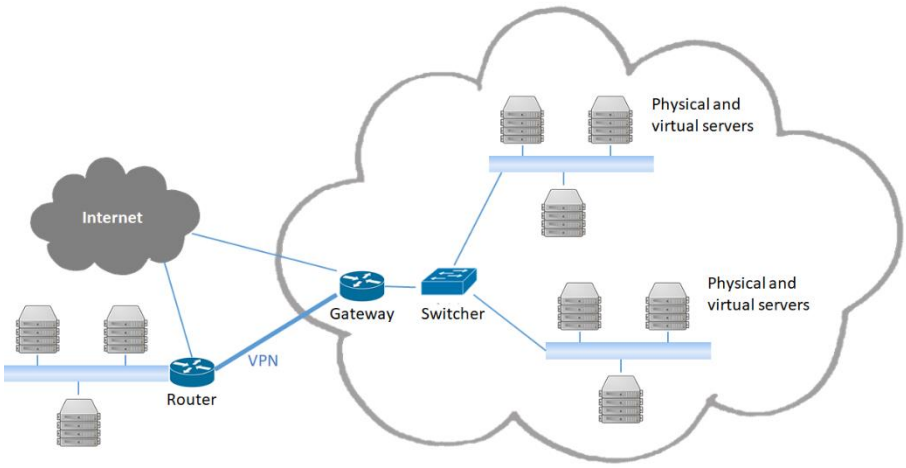

Figure 6 – Cloud as an extended enterprise data centre

Cloud platform performances depend on various cloud-services operation factors including [b-ISO/IEC 30141]:

- availability of the service;
- response time to complete service requests;

- transaction rate at which service requests are executed;
- latency for service requests;
- data throughput rate (input and output);
- number of concurrent service requests (scalability);
- capacity of data storage;
- (for IaaS and PaaS) the number of concurrent execution threads available to an application;
- (for IaaS and PaaS) the amount of random access memory available to the running program;
- data centre network Internet protocol address pool or virtual local area network range capacity.

Cloud energy consumption is influenced by its performances and, hence, its implemented services and operations.

7 End-to-end use cases addressed

For the scope of this Technical Report, three different types of cyber-physical use cases are distinguished, building on a combination of innovative technologies:

1 monitoring applications using smart IoT systems and AI software;

2 smart applications using edge computing and cloud data centres;

3 simulation applications using the digital twin pattern.

For these use case typologies, energy perspectives are analysed in relation to their computer processing and data management aspects.

7.1 Monitoring application using smart IoT systems and AI software

These applications commonly implement a three-layer deployment schema – see Figure 4.

7.2 Smart application using edge computing and cloud data centre

Smart applications and systems build on collaboration of multiple distributed smart servers that connect the physical and the digital world providing real-time data analysis and actionable intelligence – see the cyber-physical systems (clause 6.4) and the digital twin pattern (clause 6.3).

These applications commonly implement a four-layer deployment schema – see Figure 5.

7.3 Simulation applications using digital twin pattern

These applications commonly implement a four-layer deployment schema – see Figure 5.

8 Energy efficiency criteria

This clause reviews the energy efficiency criteria applicable to the use cases studied in this Technical Report. The purpose of this analysis is to give objective and quantitative energy efficiency criteria for ICT goods, networks and services used in the use cases. In the case of goods, networks and services without available quantitative energy efficiency standards, the best available technologies present in the market that can somehow potentiate the energy efficiency in the AI and emerging technologies under study are outlined.

Energy Efficiency Directive 2012/27/EU defines energy efficiency as the "ratio of output of performance, service, goods or energy, to input of energy" [b-EU D2012/27] in line with the general energy efficiency definition used in ITU-T documents, e.g., [b-ITU-T L.1330], see clause 3.1.9.

8.1 Adopted methodology

The methodology chosen to accommodate the different measures within a certain use of an ICT good, network or service, was to consider a circular value-chain process consisting of three main steps:

a data storage;

b data transfer or movement; and

c data processing or analytics.

The concept is that each of these stages of data management and use have different energy efficiency criteria and needs, and are analysed individually.

The functionalities make the emerging technologies under study, such as AI, largely depend on current ICT goods and networks, albeit with different configurations originating from the innovative aspect of these technologies. For this, the ICT goods and networks outlined in clauses 8.2 to 8.4 represent the best available practices and technologies in the field, so that this Technical Report may have an extended lifetime.

8.2 Data storage

Regarding data storage, data centres are first considered as computer warehouses that store a large amount of data for different organizations to meet their daily transaction processing needs. They contain servers for the collection of data, and network infrastructure for the utilization and storage of the data.

8.2.1 Metrics and criteria

8.2.1.1 Power usage effectiveness and data centre infrastructure efficiency

The metric commonly used by the ICT industry to identify the energy efficiency of a data centre is power usage effectiveness (PUE), which is the ratio of total data centre input power to IT load power. A higher PUE means that more energy is used by the supporting infrastructure such as lighting, cooling and power distribution, rather than energy going to IT equipment. PUE has been criticized as a measure of efficiency because it only considers energy usage and was intended only as a site-specific metric rather than one used for comparison between facilities. PUE may decrease when IT load increases because the IT equipment consumes more energy, even though the efficiency of the data centre has not actually improved [b-Brady].

The ideal value of PUE is 1.0, which indicates all energy goes to the IT equipment. However, this is generally not attainable at present due to the consumption of electricity by uninterruptible power supplies, fans, pumps, transformers, lighting and other auxiliary equipment in addition to the consuming IT load.

The most efficient data centres are approaching low values, such as EcoDataCenter in Falun, Sweden, which has a PUE of 1.15 [b-EcoDataCenter], and Google's fleet of data centres that achieved a Q2 2020 trailing 12 month global average of 1.10. However, there are indications that PUE improvements are plateauing [b-Lawrence].

PUE has been shown to correlate poorly with carbon emissions [b-Lei], so should not be the only metric tracked [b-Whitehead]. Another metric is the data centre infrastructure efficiency, E_{DCI}, which is expressed as a percentage and is calculated by dividing IT equipment power, P_{ITE}, by total facility power, P_{TF}:

$$E_{DCI} = (P_{ITE}/P_{TF}) \times 100\%$$

The EU code of conduct for data centres [b-EU JRC TR108354], a voluntary programme, has been created in response to their increasing energy consumption and the need to reduce the related environmental, economic and energy supply impacts, and with the aim to inform and stimulate operators and owners to reduce energy consumption in a cost-effective manner, and without

hampering the critical function of data centres. [b-EU JRC TR108354] gives an overview of a decrease in the average PUE of data centres throughout the years for the whole 289 data centre sample. Some best practices for improvement in the overall PUE of data centres are given [b-EU JRC TR119571][b-ITU-T L.1300], like free cooling technologies (direct and indirect air, direct and indirect water) or project management procedures that can be implemented and have a positive impact on the overall efficiency of the structures.

8.2.1.2 Energy efficiency in the power supply unit

An example of an energy efficiency criterion applicable to the stage of data storage relates to the power supply unit (PSU). Regarding the PSU, [b-EC Reg 2019/424] outlines that from 2023-01-01, for servers and online data storage products, with the exception of direct current servers and of direct current data storage products, the PSU efficiency at 10%, 20%, 50% and 100% of the rated load level and the power factor at 50 % of the rated load level shall not be less than the values reported in Table 1.

Table 1 – Minimum power supply unit efficiency and power factor requirements from 2023-01-01 [b-EC Reg 2019/424]

	Minimum PSU efficiency				Minimum Power Factor
% of rated load	10%	20%	50%	100%	50%
Multi output	-	90%	94%	91%	0,95
Single output	90%	94%	96%	91%	0,95

8.2.2 Energy efficiency best practices on data storage

A data storage product means a fully functional storage system that supplies data storage services to clients and devices attached directly or through a network [b-ISO/IEC 30141].

Both Energy Star [b-Energy Star] and the *EU code of conduct on data centre energy efficiency* [b-EU JRC TR119571] present a set of energy efficiency measures that can be implemented in data storage facilities. The concepts employed aim to make better use of existing storage hardware, reducing the volume of data to be stored and using storage equipment that consumes less energy.

To make better use of existing storage hardware, storage tiers are outlined as the assignment of different categories of data to various types of storage media, with the ultimate goal of reducing the storage cost. These tiers are determined by performance needs, the cost of the storage media and how often this data is accessed. There is a grading system for tier storage, where the most frequently accessed data are placed in the highest performing storage, while rarely accessed data go in low-performance, cheaper storage.

Storage virtualization is another measure to improve energy efficiency in the storage stage. Storage virtualization is the pooling of physical storage from multiple network storage devices into what appears to be a single storage device that is managed from a central console. Storage virtualization enhances storage performance, enables the use of storage tiers and makes it easier to expand storage capacity.

Thin provisioning presents an application with a virtual volume of just about any size, but allocates physical storage space on a just-enough, just-in-time basis by centrally controlling capacity and allocating space to an application as data is actually written.

Another big part of energy efficiency in data storage measures relies on the reduction of the volume of data to be stored. Data compression is performed by software that uses a formula or algorithm to determine how to shrink the size of the data. Compression functionality is built into a wide range of

technologies, including storage systems, databases, operating systems and software applications used by enterprise organizations.

Deduplication software works by retaining one unique instance of a file or data block and replacing all duplicates with a pointer to the original.

Snapshot technology is another measure that can be applied in order to save energy in the storage stage. Snapshot technology works so as to avoid downtime; instead of making a full backup of the data, high-availability systems may instead perform the backup on a snapshot – a read-only copy of the data set frozen at a point in time – and allow applications to continue writing to their data. Snapshots create temporary virtual copies of data that only include data changes.

Not specific related to data centre storage, but very important, are other strategies like the decommissioning of unused servers and the consolidation of lightly utilized servers. These strategies can also be measures to be implemented since data centres often possess aged and servers of no use that are still running. The management of airflow is an important aspect of data centre energy optimization. Some of the best practices include the redefinition of the server racks into a hot- or cold aisle layout, where the rows of server racks are oriented so that the fronts of the servers face each other instead of being in the same air flow direction.

The containment or enclosures arrangements when used in combination with the hot- or cold-aisle layout can also improve the efficiency of data centre server rooms. This containment refers to the various physical barriers that eliminate the mixing of cold air with hot air coming out of the server racks. This configuration allows for higher temperatures in the server rooms, which saves energy due to the slowing down of fan speeds and increase of chilled water temperatures, as well as the increase of the use of free cooling techniques.

Another measure that can be implemented within the design of data centres are variable speed fan drives for computer room air conditioning that can be adjusted on the demand of the data centre, which is constantly changing.

Finally, Energy Star indicates that proper deployment, like the positioning of diffusers, blanking panels, structured cabling systems, the elimination of sub-floor obstructions, floor grommets and the correct placing of vented tiles for airflow management devices, is a measure to improve the overall efficiency of the data centre.

8.3 Data transfer – 5G, wireless, and copper networks

This clause evaluates the best available technologies regarding data transfer and networks. The scope covers the evaluation of the network technologies and data transfer protocols aimed at better energy efficiency in the use cases studied.

Regarding mobile networks, 5G will highly contribute to the accomplishment of the expectations of the IoT ecosystem and all its interdependent stakeholders in terms of accessibility and network speeds.

Studies have found 5G to be up to 90 percent more energy efficiency per traffic unit than legacy 4G networks, with several hardware and software solutions that help to save energy [b-Nokia]. However, experts expect, similarly to what happened with other technologies before, that the deployment of 5G network will lead to an increase in energy consumption. With companies expecting to increase their energy consumption due to an increase in traffic, the technology needs to be rolled out in such a way that the higher energy consumption of this technology is met with appropriate measures that minimize this increase. Some of the measures outlined in a recent position paper [b-Ericsson] that may contribute to a smoother and more efficient technology transition are:

- prepare the network: which includes the replacement of existing networks instead of adding new ones;

- activate energy-saving software: which refers to energy saving features in 5G network components, such as micro-sleep functions;
- build 5G with precision: which concerns the avoidance of over-dimensioning hardware by considering the needs of the area of installation, to save on energy and costs;
- operate site infrastructure intelligently: including the use of AI to operate site infrastructure.

On another aspect, presently, the impact of wireless networks on the energy footprint of the ICT sector, could be said to be quite small due to the protocols that are used. Nevertheless, as more and more traffic is being transferred towards wireless networks, with the IoT being heavily dependent on wireless technologies, the traffic is also increasing, thus making the energy efficiency of such communications non-negligible.

For example, in an evaluation with four of the most popular IoT protocols (Zigbee, LoRa, Bluetooth and WiFi) that constituted a wireless sensor network, in a smart campus experience [b-Del-Valle Soto], an assessment of the energy efficiency of these protocols was performed. With a network of sensors being composed of sensors, radio transmitters and receivers, CPU and memory and the power source (battery), the authors identify some issues that can affect the node battery consumption. The term "unbalanced energy depletion" is presented and describes a situation where the nodes that are closer to the coordinator node carry more traffic, and so they consume more energy than those nodes further away from the root node. This imbalance causes the overall energy to be distributed non-uniformly in the network, causing some nodes to run out of power faster than others. Looking at the energy side of the networks, the authors have identified that this is an important issue in the warranty of stable networks, due to the life of batteries. In this network, the authors found the most efficient IoT protocol to be Zigbee, both in cooperative and collaborative configurations of the network.

8.3.1 Energy efficiency of 5G base stations

According to [b-ITU FG-AI4EE D.WG3-2], the introduction of 5G into the networks with 2G, 3G and 4G brings more power consumption. Moreover, though 5G can provide faster and more services, its energy efficiency is not always optimal especially at the initial stage of deployment or with low traffic. On the other side, it is possible to re-use 4G energy saving practices and technologies (e.g., carrier shutdown, channel shutdown and symbol shutdown) [b-ITU FG-AI4EE D.WG3-2]. While enhanced technologies have been developed in 5G era (e.g., deep sleep and symbol aggregation shutdown). Finally, BD and AI must be further utilized to implement intelligent and self-adaptive energy saving solutions and strategies, based on specific site traffic and other site-related conditions. According to [b-ITU FG-AI4EE D.WG3-2], an AI-driven smart procedure for energy saving includes the following steps.

1 Data acquisition: The network performance data and measurement report/call detail trace data of the base station are obtained through network management or data acquisition system.

2 Data processing: The collected data are pre-processed as being cleaned, constructed, aggregated and screened as training data for scene recognition, load forecasting and other models.

3 Scenario identification: The ML algorithm is used to identify the application scenario and determine the energy-saving shutdown scheme and function.

4 Threshold determination: According to the energy-saving target to be achieved, the appropriate energy-saving threshold is determined.

5 Time-span determination: Based on historical traffic data, the ML algorithm is used to predict the traffic load in a certain period of time in the future to determine the energy saving time and activate the time window.

6 Execution strategy: The integrated energy-saving strategy is sent to the network management system to perform energy-saving operations on the 5G base station, such as deep sleep, carrier shutdown, symbol shutdown and corresponding activation time window.

7 Feedback and optimization: The performance data of the base station are collected to evaluate whether the expected target is achieved and the closed-loop iterative optimization threshold strategy is adopted.

Naturally, in addition to network traffic monitoring, there are also AI techniques for traffic forecast.

8.3.2 Energy efficiency best practices with passive optical networks

Communication networks are responsible for a great amount of energy consumption in the ICT ecosystem. Passive optical networks (PONs) are important to consider when addressing energy efficiency practices. One of the most promising methods to save energy in fibre access networks is to put network devices (an optical network unit (ONU) or optical line termination or parts of them) into sleep mode when there is no traffic to be transmitted. However, putting these devices into sleep mode may incur packet delays. Ultimately, to save energy in PONs, using an efficient energy management with scheduling for the sleep and wake-up period is a challenging task, although rewarding.

Energy saving of optical networks at four different levels is addressed in [b-Zhang]: components; transmission; network; and applications.

- Component level, integrating all-optical processing components, such as optical buffers, switching fabrics and wavelength converters, may significantly reduce energy consumption.
- Transmission level, low-attenuation and low-dispersion fibres, energy-efficient optical transmitters and receivers (which improve the energy efficiency of transmission) are also being introduced.
- Network level, energy-efficient resource allocation mechanisms, green routing, long-reach optical access networks, etc. are trying to reduce energy consumption of optical networks.
- Application level, mechanisms for energy efficient network connectivity such as proxying and green approaches for cloud computing are being proposed to reduce energy consumption.

[b-Valcarenghi] describes several solutions to reduce energy consumption in ONUs that have been proposed by many researchers. The article reports that for time-division multiplexed PONs in sleep mode, ONU energy consumption can be reduced by switching them to low power mode when idle. However, huge savings can only be achieved if ONUs are capable of quickly regaining synchronization on wake-up, and the power consumed while sleeping is much less than when the ONU is on. Moreover, data link layer solutions alone (e.g., sleep mode) may be effective when network utilization is low, but when network utilization increases, physical layer support (e.g., quick resynchronization) is necessary. If solutions and ONU architectures with such characteristics are developed, huge energy savings and limited delay increase can be achieved.

8.3.3 Energy efficiency best practices with copper networks

The ecological footprint of broadband access technology has substantially increased during the last decades due to the rapid acceptance and availability of fast and reliable Internet connections. At the same time, an increase of energy consumption in these networks has also been detected. Despite the fact that there is a clear trend to more fibre-based access solutions in the future, access based on digital subscriber lines (DSLs) will remain relevant.

There are several initiatives, like the *EU code of conduct on energy consumption of broadband equipment* [b-EU JRC TR106039] and others that have been addressing the issue of energy consumption and urging the ICT industry to act on its environmental impact.

DSL is a cost-effective solution to bring broadband access to its customers by using the existing copper infrastructure, originally installed for simple voice communication.

[b-Guenach] on improving the energy efficiency of broadband copper access networks outlines some conclusions on best practices to be implemented in broadband networks.

"… at the architecture level, by moving to smaller nodes, and at the component level, by selecting energy-efficient technology … and/or by introducing design techniques (for instance, clock and power gating), which benefit from the typical Internet usage (burstiness of traffic, video streaming, day/night cycles), power savings can be obtained, which, in the best case, can reduce by more than half the energy consumption of the access network. … moving to a more distributed access network of small nodes is currently necessary with the introduction of G.vector and the planned introduction of G.fast, one can rethink how these nodes can be cooled (fresh air cooling, passive cooling) and powered (reverse powering) because of the different scales of these nodes. Some of these solutions not only provide energy saving but also assist in enabling the operator to easily deploy these increased numbers of small-sized active components in this network. For instance, by removing the need of active cooling (requiring additional power) in the small nodes, the power budget to feed these small nodes comes within reach of new power schemes such as reverse powering (which do introduce some inefficiency compared with classical powering schemes such as local ac/dc conversion)." p. 576 of [b-Guenach]

The EU code of conduct outlines also that

"The volume of deployed broadband equipment is increasing dramatically and so does its combined power consumption. Due to low customer aggregation ratios (typically, one CPE [*customer premises equipment*] per customer), such equipment is typically idle most of the time, most of the time exchanging data only to maintain its network status. It is therefore evident that such equipment can be optimized in terms of its power consumption and activity profiles. Examples of such techniques include dynamic adaptation (e.g., performance scaling), smart standby (e.g., through proxying network presence and virtualization of functions) and energy aware management." p. 56 of [b-EU JRC TR106039]

8.4 Data processing

On data processing, there are several groups of ICT goods that can be evaluated. At first, computers and small servers are the ones that may present as more obvious, since these will be responsible for a great part of data processing with the roll-out of the emerging technologies, with the data being processed in cloud or edge computing.

[b-EC Reg 2019/424] outlines some guidelines for desktop computers, integrated desktop computers, notebook computers (including tablet computers, slate computers and mobile thin clients), desktop thin clients, workstations, mobile workstations, small-scale servers and computer servers. These types of hardware are ultimately expected to be responsible for a big part of the processing of data within the IoT environment. The Ecodesign requirements are presented in Annex II of the Guidelines.

8.4.1 Energy efficiency best practices on data processing

There are several cases of energy efficiency in the different parts of the life of a byte of data that have been studied, especially in terms of the hardware or the infrastructures being used. An area that is still less investigated is the role that coding and software can have in the energy performance in the data processing stages.

A study on the empirical evaluation of two best practices for energy-efficient software development [b-Procaccianti] evaluated the impact of two best practices for energy-efficient software and applied these practices in two widely used software applications, MySQL Server and Apache Webserver and each practice successfully reduced the energy consumption of our test environment and conclude that software design and implementation choices significantly affect energy efficiency. This study is based on previous studies that have been evaluating the connection between software development and energy efficiency, like the work of [b-Capra], which analyses the impact of application development environments over the energy efficiency of software applications, concluding for example that a high framework entropy is beneficial for the energy efficiency of small and medium applications. The work of [b-Sahin], which investigates the energy impact of using software design patterns, concludes that the impact of applying a design pattern varies greatly, from less than 1% to more than 700%. The work of [b-Noureddine] analyses the energy impact of programming languages and algorithmic choices, and finds that the algorithm choice has a significant impact on energy consumption (a

recursive algorithm is more energy-efficient than an iterative one) and that the chosen programming language also has a significant impact on energy consumption. [b-Manotas] investigated the energy impact of web servers in web applications and found that the energy consumption of a web application greatly varies depending on the chosen web server and that the variation depends on the specific feature of the web server. The same web server might be more energy efficient in a specific scenario (e.g., search) and very inefficient in others.

[b-Song], which analyses energy efficiency optimization in a BD processing platform by improving resources utilization, outlines a proposal for resources utilization in BD processing by allocating different resources according to a task-related best resource ratio (BRR), such as "CPU, disk, network in the ratio of 1:2:4", rather than the resource's quantity, such as "CPU = 1 GHz, network = 20 MB/s". The study deduces the BRR of data processing tasks, and designs a resource ratio based approach (R^2), which includes a task scheduling algorithm and resource allocation algorithm, for energy efficiency optimization. Experiments show that the R^2 approach can improve energy efficiency by 10%.

In 2011, Intel published a white paper on energy-efficient software guidelines, which can be help developers aiming to reduce the energy consumption of their pieces of software [b-Intel].

8.4.2 Machine learning energy consumption and efficiency

Power models are built to design better hardware, design better algorithms or design better software to map these algorithms on to hardware.

In the case of ML applications, at the system-level, it is possible to distinguish between two power estimation models [b-García-Martín] as follows.

- Software level: The focus is on the energy consumption of the application or software implementation and exploration of optimization techniques, working at the level of: application (e.g., kernel sizes in a neural network); instructions (e.g., by using performance counter profiling, understanding the cost for each instruction and trying to reduce the most expensive part of the code).
- Hardware-level: The focus is on the energy consumption of specific hardware components (e.g., processor, memory and input or output peripherals).

A general survey of the techniques utilized for both software and hardware levels is provided in [b-García-Martín].

In ML, two main phases can be further distinguished: training and inference (or operational). Research settings typically focus on model training and accuracy performance. In industrial settings, the cost of inference might exceed the training costs in the long term. In this context, it might be more beneficial to use more expensive models to train even if they are more efficient in the inference phase, when in operation.

Motivation: In the context of deep neural networks regarding the training of some ML algorithms, the process generally involves a certain number of passes through the dataset, often called epochs. It has been realized in a recent study that there is a threshold of epochs at which the accuracy of ML model reaches a plateau [b-Thompson], but energy consumption continues to increase [b-Strubell]. The same applies to larger training data sets that demand more energy to train, but do not lead necessarily to a proportional benefit in accuracy. The study suggests that there may be a path for models not reaching full accuracy and still complying with the needs of the user. Another suggestion is to use transfer learning where an existing model could be repurposed for a different task in order to save energy and time.

Advances in techniques and hardware for training deep neural networks have recently enabled impressive accuracy improvements in image processing and across many fundamental natural language processing (NLP) tasks as reported in [b-Strubell] with also a heavy dependence on large computational resources that need similarly substantial energy consumption. This clause mainly

focuses on deep learning algorithms, as it is the set of techniques currently creating most power consumption models. However, it is worth noting that this is just a subset of ML algorithms, and practitioners should also be encouraged to resort to more traditional and power-efficient options (e.g., random forest or XGBoost) when appropriate.

Some of the actionable recommendations to reduce costs in NLP that could be adapted in other applications mentioned in [b-Strubell] include:

- the reporting of training time and sensitivity to hyperparameters where it would be beneficial to directly compare different models to perform a cost-benefit (accuracy) analysis;
- prioritize computationally efficient hardware and algorithms.

Most recently, some research developed within Google has also raised awareness of the high computational needs and other ethical risks of the latest NLP models [b-Bender], which has put some pressure on Google researchers to emphasize the importance of the topic and justify how the benefits of the model outweigh the energy costs. Another study [b-Patterson] shows how, depending on the choices made for training a large NLP model, like the type of ML model, data centre and processor; carbon footprint can be reduced by about 100 to 1 000 times.

How to estimate emissions

The research community is more and more asking for a systematic way of reporting the energy and carbon footprints of ML models [b-Henderson], and also calling to direct research towards more efficient models focusing on what is called green AI, to decrease its carbon footprint and increase its inclusivity [b-Schwartz]. On the other side of the balance there is so-called red AI, which is AI that targets accuracy using massive computational power.

There are already some efforts to reduce the size of these models through techniques like distillation or quantization [b-Hinton][b-Zafrir]. These efforts still, however, rely on a significant amount of processing to produce these reductions. Sparsely activated models have claimed to provide significant savings both for training and inference [b-Fedus].

In recent years, some tools, like the green algorithms tool [b-Fedus] or ML emissions calculator [b-Lacoste], have been created that help ML researchers and practitioners get an understanding of the approximate environmental impact of their experiments. They based their estimations on the location of the server used, the length of the procedure, and the make and model of the hardware. [b-Bender] presents state-of-the-art approaches and software tools to estimate energy consumption from ML algorithms up to mid-2019.

However, being able to approximately measuring while running the experiments is generally more accurate than indirect estimations a posteriori. At the end of 2019, [b-Lottick] presented a python package that calculates the energy and CO_2 emissions of any (python) function and provides an energy usage report to add context to these results (see Figure 7). Another attempt presented a month later is the experiment-impact-tracker framework [b-Henderson], which provides a systematic way for the community to consistently account and report energy, computation and carbon metrics. This tool can make it easier to understand the full training lifecycle, including training attempts performed before the setup used is the final one. Most recently CodeCarbon [b-mlco2/codecarbon] is also being developed to approximately measure carbon emissions in a more automatic way.

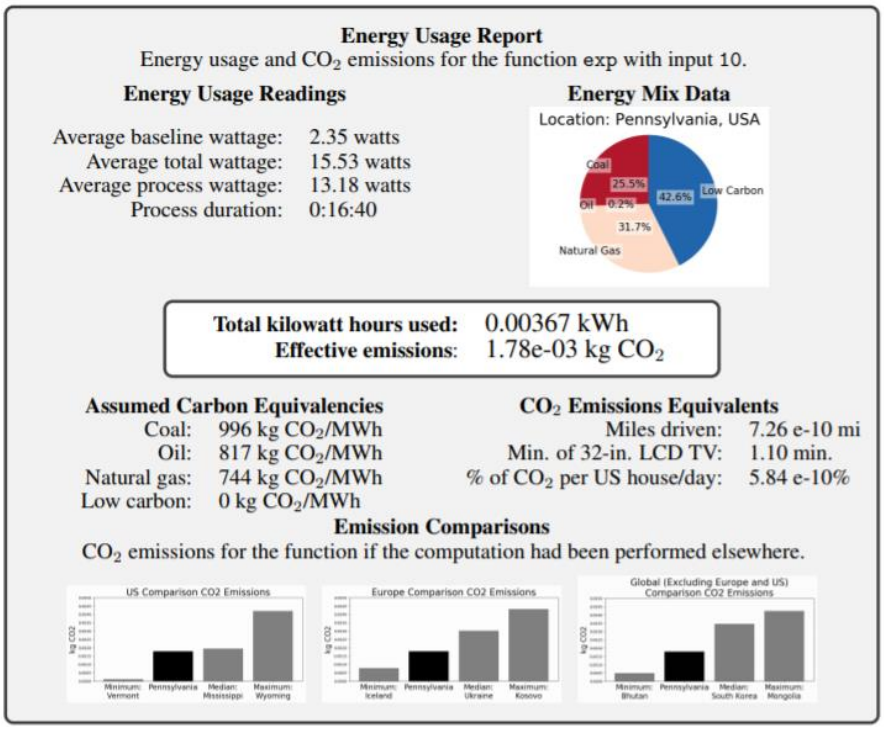

Figure 7 – An example energy usage report for a simple exponential function. Image taken directly from [b-Lottick]

Machine learning on edge devices

More and more computation is gradually taking place on edge devices (i.e., IoT), thanks to the explosive growth of Internet-connected devices. Edge devices are power or resource constrained, and typically, a tradeoff between accuracy and efficiency must be found. [b-Kumar] considers new hardware architecture for ML on edge and hardware-based full stack optimization for ML on edge computing as two categories that can be exploited in order to meet the growing demand for resources in ML algorithms.

The same applies at the software level and packages, where now many projects focus on developing and optimizing software platforms and ML packages to meet the low-power requirements of the edge. Algorithms, again, are seen as a key path to reducing the energy consumption of ML models, by reducing the computational requirements and the accuracy of operations and operands.

In order to address the issue of the large environmental impact of such AI training processes, some solutions have been outlined [b-Cai]. A "once-for-all" network trains a large model that has many pretrained sub-models of different sizes that can be tailored to a range of platforms without retraining.

Each of these sub-models can operate independently at inference time without retraining, and the system identifies the best sub-model based on the accuracy and latency trade-offs that correlate to the target hardware's power and speed limits. See Figure 8.

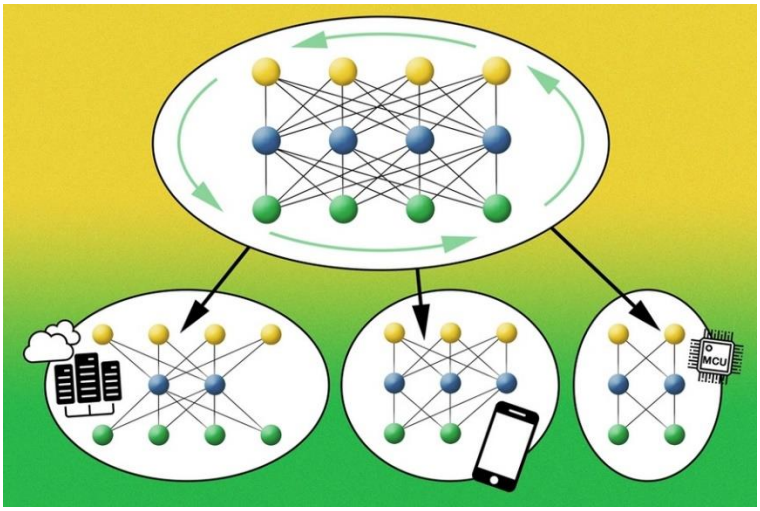

Figure 8 – Neural network with sub-models. Source: MIT (Source: [b-Matheson])

A "progressive shrinking" algorithm trains the large model to support the sub-models at the same time. First, the large model is trained and then the smaller sub-models are trained with the aid of the large model so that they learn simultaneously. Finally, all sub-models are supported, allowing for a speedy specialization based on the specifications of the target platform.

Summary

In summary, most of the latest research on the topic recommends the:

– use of sparsely activated deep neural networks for energy savings;

– payment of attention to the geographic location of the servers where ML workload runs;

– use of specialized infrastructure that includes accelerators appropriate for the task;

– duereporting of the energy consumption and CO_2 emissions on ML papers and research, especially when it involves large training of models;

– use of efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics;

– inclusion of the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly;

– taking energy needs for inference into account, as these can often outweigh the training ones; and

– release of pre-trained models to save others the cost of retraining them.

This list is non-exhaustive but should serve as an indication of the state of play at the time of writing. There are also experts who are urging policymakers to stimulate transparency and the creation of standards, to facilitate the emission calculations in the area of AI [b-Dhar].

9 Appliance of energy criteria to the end-to-end considered by this Technical Report

The energy efficiency good practices (introduced in clause 8) are applied to the three end-to-end application typologies introduced in clause 7, acknowledging the different components and the three main steps of the circular value chain discussed.

9.1 Monitoring application using smart IoT systems and AI software

In the case of developing and managing monitoring applications by using smart IoT systems and AI software, good practices for energy efficiency are summarized in Table 2.

Table 2 – Energy efficiency good practices for implementing monitoring application using smart IoT systems and AI software

Component	Value-chain step to which it contributes	Energy efficiency good practices	Notes
	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics,
			See clause 8.4.2

Table 2 – Energy efficiency good practices for implementing monitoring application using smart IoT systems and AI software

Component	Value-chain step to which it contributes	Energy efficiency good practices	Notes
		• include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly, • keep energy needs for inference into account, as these can often outweigh the training ones and • release pre-trained models to save others the cost of retraining them.	
c	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization ;
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point
			See clause 8.4.2

Table 2 – Energy efficiency good practices for implementing monitoring application using smart IoT systems and AI software

Component	Value-chain step to which it contributes	Energy efficiency good practices	Notes
		operations), in combination with accuracy and other similar metrics, • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly, • keep energy needs for inference into account, as these can often outweigh the training ones and • release pre-trained models to save others the cost of retraining them.	

9.2 Smart application using edge computing and cloud data centre

In the case of developing and managing smart applications building on edge computing and a cloud data centre, good practices for energy efficiency are summarized in Table 3.

Table 3 – Energy efficiency good practices for implementing smart application using edge computing and cloud data cent**re**

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy;	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

Table 3 – Energy efficiency good practices for implementing smart application using edge computing and cloud data cent**re**

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• feedback and optimization.	

	Data processing (ML)	One or more of the following techniques/strategies:
		• Hardware-based full stack optimization;
		• A "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics,
		  • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly,
		  • keep energy needs for inference into account, as these can often outweigh the training ones and
		  • release pre-trained models to save others the cost of retraining them.
			See clause 8.4.2

	Data storage	One or more of the following techniques: • storage tiers;	See clause 8.2

Table 3 – Energy efficiency good practices for implementing smart application using edge computing and cloud data cent**re**

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • Scenario Identification; • Threshold Determination; • Time-span Determination; • Execution Strategy; • Feedback and Optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)
	Data processing (ML)	One or more of the following techniques/strategies: • Hardware-based full stack optimization • A "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably: • use sparsely activated deep neural networks for energy savings; • pay attention to the geographic location of the servers where ML workload runs; • use of specialized infrastructure that includes accelerators appropriate for the task; • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models; • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics; • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly;	See clause 8.4.2

Table 3 – Energy efficiency good practices for implementing smart application using edge computing and cloud data cent**re**

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• keep energy needs for inference into account, as these can often outweigh the training ones; and • release pre-trained models to save others the cost of retraining them.	
c	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization;
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine
			See clause 8.4.2

Table 3 – Energy efficiency good practices for implementing smart application using edge computing and cloud data cent**re**

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		learning papers and research – especially when it involves large training of models, • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics, • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly, • keep energy needs for inference into account, as these can often outweigh the training ones and • release pre-trained models to save others the cost of retraining them.	

9.3 Simulation applications using digital twin pattern

In the case of developing and managing simulation applications by applying the digital twin pattern, good practices for energy efficiency are summarized in Table 4.

Table 4 – Energy efficiency good practices for implementing simulation applications applying the digital twin pattern

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination;	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2

Table 4 – Energy efficiency good practices for implementing simulation applications applying the digital twin pattern

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• time-span determination; • execution strategy; • feedback and optimization.	and 8.3.3, respectively)
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization;
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics,
		  • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly,
		  • keep energy needs for inference into account, as these can often outweigh the training ones and
		  • release pre-trained models to save others the cost of retraining them.
			See clause 8.4.2

	Data storage	One or more of the following techniques: • storage tiers;	See clause 8.2

Table 4 – Energy efficiency good practices for implementing simulation applications applying the digital twin pattern

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization;
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics,
		  • include the full training lifecycle in the calculations, which considers previous attempts needed until everything is set up correctly,
		  • keep energy needs for inference into account, as these can often outweigh the training ones and
			See clause 8.4.2

Table 4 – Energy efficiency good practices for implementing simulation applications applying the digital twin pattern

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		• release pre-trained models to save others the cost of retraining them.	
	Data storage	One or more of the following techniques: • storage tiers; • storage virtualization; • thin provisioning; • data compression; • deduplication; • decommissioning of non-used storage; • snapshot technology.	See clause 8.2
	Data transfer (5G)	Implement the AI-driven smart procedure consisting of the steps: • scenario identification; • threshold determination; • time-span determination; • execution strategy; • feedback and optimization.	See clause 8.3.1 (for optical and copper networks, see clauses 8.3.2 and 8.3.3, respectively)

	Data processing (ML)	One or more of the following techniques/strategies:
		• hardware-based full stack optimization;
		• a "once-for-all" AI network (which trains a large model that has many pretrained sub-models) – notably:
		  • use sparsely activated deep neural networks for energy savings,
		  • pay attention to the geographic location of the servers where ML workload runs,
		  • use of specialized infrastructure that includes accelerators appropriate for the task,
		  • duly report the energy consumption and CO_2 emissions on machine learning papers and research – especially when it involves large training of models,
		  • use also efficiency as an evaluation metric (e.g., floating point operations), in combination with accuracy and other similar metrics,
		  • include the full training lifecycle in the calculations, which considers
			See clause 8.4.2

Table 4 – Energy efficiency good practices for implementing simulation applications applying the digital twin pattern

Component	Value-chain steps to which it contributes	Energy efficiency good practices	Notes
		previous attempts needed until everything is set up correctly, • keep energy needs for inference into account, as these can often outweigh the training ones and • release pre-trained models to save others the cost of retraining them.	

Bibliography

[b-ITU-T L.1300] Recommendation ITU-T L.1300 (2015), *Best practices for green data centres*.

[b-ITU-T L.1330] Recommendation ITU-T L.1330 (2015), *Energy efficiency measurement and metrics for telecommunication networks*.

[b-ITU-T Y.2221] Recommendation ITU-T Y.4105/ITU-T Y.2221 (2010), *Requirements for support of ubiquitous sensor network (USN) applications and services in the NGN environment*.

[b-ITU-T Y.3502] Recommendation ITU-T Y.3502 (2014), *Information technology – Cloud computing – Reference architecture*.

[b-ITU-T Y.3600] Recommendation ITU-T Y.3600 (2015), *Big data – Cloud computing based requirements and capabilities.*

[b-ITU-T Y.4051] Recommendation ITU-T Y.4051 (2019), *Vocabulary for smart cities and communities*.

[b-ITU FG-AI4EE D.WG3-2] ITU FG-AI4EE D.WG3-2 (2021), *Smart energy saving of 5G base station: Based on AI and other emerging technologies to forecast and optimize the management of 5G wireless network energy consumption*. Available [2021-11-09] at: https://www.itu.int/pub/publications.aspx?lang=en&parent=T-FG-AI4EE-2021-D.WG3.02

[b-ISO 5127] ISO 5127, *Information and documentation – Foundation and vocabulary.*

[b-ISO/IEC 20924] ISO/IEC 20924:2021, *Information technology – Internet of Things (IoT) – Vocabulary.*

[b-ISO/IEC TR 23188] ISO/IEC TR 23188:2020, *Information technology – Cloud computing – Edge computing landscape.*

[b-ISO/TR 24464] ISO/TR 24464 :2020, *Automation systems and integration – Industrial data – Visualization elements of digital twins*.

[b-ISO/IEC/IEEE 24765] ISO/IEC/IEEE 24765:2017, *Systems and software engineering – Vocabulary*.

[b-ISO/IEC TR 29119-11] ISO/IEC TR 29119-11:2020, *Software and systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems*.

[b-ISO/IEC 30141] ISO/IEC 30141:2018, *Internet of things (loT) – Reference architecture*.

[b-EC EGAI] European Commission (2020). *Ethics guidelines for trustworthy AI.* Brussels: European Commission. Available [2011-11-02] at: https://wayback.archive-it.org/12090/20201227221227/https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[b-EC Ecodesign GL] European Commission (2014). *Guidelines accompanying Commission Regulation (EU) 617/213 implementing Directive 2009/125/EC of the European Parliament and of the Council in regards to ecodesign requirements for computers and computer servers.* Brussels: European Commission. 13 pp. Available [viewed 2021-11-03] at: https://ec.europa.eu/energy/sites/ener/files/documents/comm_2013-617_imp_2009-125_directive.pdf

[b-EC Reg 2019/424] European Commission (2019). Commission Regulation (EU) 2019/424 of 15 March 2019 laying down ecodesign requirements for servers and data storage products pursuant to Directive 2009/125/EC of the European Parliament and of the Council and amending Commission Regulation (EU) No 617/2013. *Off J Eur Union*. **L74**, pp. 46-66. Available [viewed 2021-11-03] at: https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1553786820621&uri=CELEX%3A32019R0424.

[b-EC SFC] European Commission (2020). *2020 Strategic foresight report: Strategic foresight – Charting the course towards a more resilient Europe*. Brussels: European Commission. 42 pp. Available [viewed 2021-11-02] at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0493&from=EN

[b-EU cloud market] Montevecchi, F., Stickler, T., Hintemann, R., Hinterholzer, S. (2020). *Energy-efficient cloud computing technologies and policies for an eco-friendly cloud market: Final study report*. Luxembourg: Publications Office of the European Union. 287 pp. Available [viewed 2021-11-04] from: http://op.europa.eu/en/publication-detail/-/publication/bf276684-32bd-11eb-b27b-01aa75ed71a1/language-en

[b-EU D2012/27] European Union (2012). Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency. *Off. J. Eur. Union*, **L315**, pp. 1-56. Available [viewed 2021-11-11] at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32012L0027&from=EN.

[b-EU JRC TR106039] Bertoldi, P. (2017). *EU code of conduct on energy consumption of broadband equipment*, Version 6, EUR 28519 EN, JRC Technical Report, JRC106039. Available [viewed 2021-11-08] from: https://op.europa.eu/en/publication-detail/-/publication/700037d4-150d-11e7-808e-01aa75ed71a1/language-en

[b-EU JRC TR108354] Bertoldi, P., Avgerinou, M., Castellazzi, L (2017). *Trends in data centre energy consumption under the European code of conduct for data centre energy efficiency – EUR 28874 EN*, JRC Technical Report 108354. Luxembourg: Publications Office of the European Union. 43 pp. Available [viewed 2021-11-03] at: https://publications.jrc.ec.europa.eu/repository/bitstream/JRC108354/kjna28874enn.pdf

[b-EU JRC TR119571] Acton, M. Bertoldi P., Booth J. (2021). 2021 *Best practice guidelines for the EU code of conduct on data centre energy efficiency*, version 12.1.0 (final version), JRC Technical Report 119571. Ispra: European Commission, . Available [viewed 2021-11-04] at: https://e3p.jrc.ec.europa.eu/sites/default/files/documents/publications/jrc123653_jrc119571_2021_best_practice_guidelines_final_v1_1.pdf

[b-EU JRC TR124168] Nativi, S., Craglia, M. (2021). *Destination Earth: Ecosystem architecture description, EUR 30646 EN*, JRC Technical Report 124168. Luxembourg: Publications Office of the European Union. 83 pp. Available [viewed 2021-11-03] from: https://op.europa.eu/de/publication-detail/-/publication/78e15712-8c53-11eb-b85c-01aa75ed71a1/language-en/format-PDF/source-197616568

[b-NIST FCPS] Cyber Physical Systems Public Working Group (2016). *Framework for cyber-physical systems*, Release 1.0. Gaithersburg, MD: National Institute of Standards and Technology. 266 pp.

[b-NIST SP 800-145] Mell, P., Grance, T. (2011). *The NIST definition of cloud computing: Recommendations of the National Institute of Standards and*

Technology, Special Publication 800-145. Gaithersburg, MD: National Institute of Standards and Technology. 3 pp.

[b-Adams] Adams, J., Cser, A. (2017). *Control the cloud, before the cloud controls you – Forrester data: Cloud security solutions forecast, 2016 to 2021 (global)*. Mumbai: Tata Communications. 20 pp. Available [viewed 2021-11-10] at: https://www.tatacommunications.com/wp-content/uploads/2019/02/Forrester-Report.pdf

[b-Bender] Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In: *FAccT: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610-623. Available [viewed 2021-11-09] at: http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf

[b-Bradford] Bradford, A. (2020). *The Brussels effect: How the European Union rules the world*. New York, NY: Oxford University Press. 404 pp.

[b-Brady] Brady, G.A., Kapur, N., Summers, J.L., Thompson, H.M. (2013). A case study and critical assessment in calculating power usage effectiveness for a data centre. *Energy Convers. Manage.* **76**, pp. 155–161. doi:10.1016/j.enconman.2013.07.035.

[b-Cai] Cai, H., Gan, C., Wang, T., Zhang, Z., Han, S. (2020). Once-for-all: Train one network and specialize it for efficient deployment. Poster presented at: *International Conference on Learning Representations*, 2020. Available viewed [2021-11-09] at: https://openreview.net/forum?id=HylxE1HKwS.

[b-Capra] Capra, E., Francalanci, C., Slaughter, S.A. (2012). Is software green? Application development environments and energy efficiency in open source applications. *Inf. Softw. Technol.* **54**, 60–71.

[b-Cisco] Cisco (2020). *Cisco annual Internet report (2018–2023) white paper*. San Jose, CA: Cisco. Available [viewed 2021-11-10] at: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

[b- mlco2/codecarbon] *mlco2/codecarbon* (2021). San Francisco, CA: Github. Available [viewed 2021-11-10] at: https://github.com/mlco2/codecarbon

[b-Del-Valle Soto] Del-Valle Soto, C., Valdivia, L.J., Velázquez, R., Rizo-Dominguez, L., López-Pimentel, J.-C. (2019). Smart campus: An experimental performance comparison of collaborative and cooperative schemes for wireless sensor network. *Energies* **12**, 3135. Available [viewed 2021-11-03] from: https://www.mdpi.com/1996-1073/12/16/3135

[b-Dhar] Dhar, P. (2020). The carbon impact of artificial intelligence. *Nat Mach Intell*, **2**, pp. 423–425. https://doi.org/10.1038/s42256-020-0219-9

[b-EcoDataCenter] EcoDataCenter (2021). *The world's first climate positive data center*. Available [viewed 2021-11-04] at: https://ecodatacenter.se/

[b-Energy Star] Energy Star (Internet). *Implement efficient data storage measures*. Washington, DC: Environmental Protection Agency. Available [viewed 2021-11-03] at: https://www.energystar.gov/products/implement_efficient_data_storage_measures

[b-Ericsson] Ericsson (2020). *Breaking the energy curve: How to roll out 5G without increasing energy consumption*. Stockholm: Ericsson.

Available [viewed 2021-11-04] from: https://www.ericsson.com/en/news/2020/3/breaking-the-energy-curve

[b-Fedus] Fedus, W., Zoph, B., Shazeer, N. (2021). *Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity*: https://arxiv.org/abs/2101.03961.

[b-Ferreboeuf] Ferreboeuf, H., editor (2019). *Lean ICT: Towards digital sobriety*. Paris: Shift Project. 90 pp. Available [viewed 2011-11-02] at: https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Report_The-Shift-Project_2019.pdf

[b-Flexera] Flexera (2019). *RightScale 2019 state of the cloud report*. Itasca, IL: Flexera. 50 pp.Available [viewed 2021-11-10] at: https://resources.flexera.com/web/media/documents/rightscale-2019-state-of-the-cloud-report-from-flexera.pdf

[b-García-Martín] García-Martín, E., Faviola Rodrigues, C., Graham Riley, G., Håkan Grahn, H., 2019, Estimation of energy consumption in machine learning, *J. Parallel Distrib. Comput*. **134**, pp. 75–88. Available [viewed 2021-11-04] at: https://reader.elsevier.com/reader/sd/pii/S0743731518308773?token=1E96411B916ADCA9587647E629B911B684C890BA5FB1BF38A9D74DD9AE8A90142F7A14F24A7B718D0AD2D23B64DB1BCF&originRegion=eu-west-1&originCreation=20211104082941

[b-Google] Google (Internet). *Our infrastructure*.. Mountain View, CA: Google. Available [viewed 2021-11-10] at: https://peering.google.com/#/infrastructure

[b-Greenpeace] Greenpeace East Asia, North China Electric Power University (2019). *Powering the cloud: How China's Internet industry can shift to renewable energy*. Beijing: Greenpeace. 12 pp. Available [viewed 2021-11-03] at: https://www.greenpeace.org/static/planet4-eastasia-stateless/2019/11/7bfe9069-7bfe9069-powering-the-cloud-_-english-briefing.pdf.

[b-Guenach] Guenach, M., Ben Ghorbel, M., Hooghe, K. (2015). Improving the energy efficiency of broadband copper access networks: Review and performance analysis. *IEEE Systems Journal*. **11**, pp. 562-577. doi: 10.1109/JSYST.2015.2437413.

[b-Henderson] Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., Pineau, J. (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *J. Mach. Learn. Res.* **21(**248), pp. 1–43. Available [viewed 2021-11-03] at: http://jmlr.org/papers/v21/20-312.htm

[b-Hinterman] Hinterman R. Clausen J. (2016). Green cloud? The current and future development of energy consumption by data centers, networks and end-user devices. In: *Proceedings of ICT for sustainability 2016*, pp. 109-115. Dordrecht: Atlantis Press. Available [viewed 2021-11-03] from: https://www.atlantis-press.com/proceedings/ict4s-16/25860373

[b-Hinton] Hinton, G., Vinyals, O., Dean, J. (2015). *Distilling the knowledge in a neural network*. arXiv preprint arXiv:1503.02531.

[b-Intel] Intel Software Solutions Group, Larsson, P. (2011). *Energy-efficient software guidelines*, white paper. Santa Clara, CA: Intel Corporation. 12 pp. Available [viewed 2021-11-04] at: https://software.intel.com/content/dam/develop/external/us/en/documents/energy-efficient-software-guidelines-v3-4-10-11-140545.pdf

[b-Kumar] Kumar, M., Zhang, X., Liu, L., Y. Wang, Y., Shi, W. (2020). Energy-efficient machine learning on the edges. In: *2020 IEEE International Parallel and Distributed Processing Symposium Workshops*

(IPDPSW), New Orleans, LA, USA, pp. 912–921, doi: 10.1109/IPDPSW50202.2020.00153.

[b-Lacoste] Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T. (2019). *Quantifying the carbon emissions of machine learning*: https://arxiv.org/abs/1910.09700

[b-Lawrence] Lawrence, A. (2020). *Data center PUEs flat since 2013*. Seattle, WA: Uptime Institute. Available [viewed 2021-11-04] at: https://journal.uptimeinstitute.com/data-center-pues-flat-since-2013/

[b-Lei] Lei, N., Masanet, E. (2020). Statistical analysis for predicting location-specific data center PUE and its improvement potential. *Energy*. **201**, 117556. Available [viewed 2021-11-04] at: https://reader.elsevier.com/reader/sd/pii/S0360544220306630?token=2816CD794D9EBBA2C3CD8C7BB12F65FD44227C081A6BCC06F1733BCF509AAA569D13641EA8AB90E7B7D56DC22967B877&originRegion=eu-west-1&originCreation=20211104060250.

[b-Leiserson] Leiserson, C.E., Thompson, N.C., Emer, J.S., Kuszmaul, B.C., Lampson, B.W., Sanchez, D., Schardl, T.B. (2020). There's plenty of room at the Top: What will drive computer performance after Moore's law? *Science*. **368**, eaam9744. Available [viewed 2021-11-03] at: https://doi.org/10.1126/science.aam9744

[b-Lottick] Lottick K., Susai, S., Friedler, S.A., Wilson, J.P. (2019). *Energy usage reports: Environmental awareness as part of algorithmic accountability*. https://arxiv.org/abs/1911.08354

[b-Manotas] Manotas, I., Sahin, C., Clause, J., Pollock, L., Winbladh, K. (2013). Investigating the impacts of web servers on web application energy usage. In: *2013 2nd International Workshop on Green and Sustainable Software* (GREENS), pp. 16-23, doi: 10.1109/GREENS.2013.6606417

[b-Masanet] Masanet, E., Shehabi, A., Lei, N., Smith, S., Koomey, J. (2020), Recalibrating global data center energy-use estimates. *Science*. **367**, pp. 984–986. Available [viewed 2021-11-03] at: doi:10.1126/science.aba3758.

[b-Matheson] Matheson, R. (2021). Reducing the carbon footprint of artificial intelligence: MIT system cuts the energy required for training and running neural networks., *MIT News*: Available [viewed 2021-11-03] at: https://news.mit.edu/2020/artificial-intelligence-ai-carbon-footprint-0423

[b-Mytton, 2020a] Mytton, D. (2020a). Assessing the suitability of the Greenhouse Gas Protocol for calculation of emissions from public cloud computing workloads. *J. Cloud Comput.* **9**, pp. 45. Available [viewed 2021-11-03] at: doi: 10.1186/s13677-020-00185-8.

[b-Mytton, 2020b] Mytton, D. (2020b). How much energy do data centers use? Available [viewed 2021-11-03] at: https://davidmytton.blog/how-much-energy-do-data-centers-use/

[b-Nokia] Nokia (2020). *Nokia confirms 5G as 90 percent more energy efficient* [press release]. Espoo: Nokia. Available [viewed 2021-12-28] at: https://www.nokia.com/about-us/news/releases/2020/12/02/nokia-confirms-5g-as-90-percent-more-energy-efficient/

[b-Noureddine] Noureddine, A., Bourdon, A., Rouvoy, R., Seinturier, L. (2012). A preliminary study of the impact of software engineering on green IT. In: *2012 First International Workshop on Green and Sustainable Software* (GREENS), 2012, pp. 21-27, doi: 10.1109/GREENS.2012.6224251

[b-Patterson] Patterson, D., Gonzalez, J. Le, Q., Liang, C., Munguia , L.-M., Rothchild, D., So, D., Texier, M., Dean, J. (Internet). *Carbon emissions and large neural network training*. https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf

[b-Procaccianti] Procaccianti, G., Fernandez, H., Lago, P.. (2016). Empirical evaluation of two best practices for energy-efficient software development. *J. Syst. Softw.* **117**, pp. 185–198- doi: 10.1016/j.jss.2016.02.035.

[b-Sahin] Sahin, C., Cayci, F., Gutierrez, I.L.M., Clause, J., Kiamilev, F., Pollock, L., Winbladh, K. (2012). Initial explorations on design pattern energy usage. In: *2012 First International Workshop on Green and Sustainable Software* (GREENS), pp. 55-61, doi: 10.1109/GREENS.2012.6224257

[b-Sayadi] Sayadi, B., Roullet, L. (2018). 5G: Platform and not protocol. *IEEE Softwarization*. Available [viewed 2022-01-09] at: https://sdn.ieee.org/newsletter/january-2018/5g-platform-and-not-protocol

[b-Schwartz] Schwartz, R., Dodge, J., Smith, N.A., Etzioni, O. (2020). Green AI. *Commun. ACM* **63**(12), pp. 54–63. https://dl.acm.org/doi/10.1145/3381831

[b-Shebabi, 2016] Shehabi, A., Smith, S.J., Horner, N., Azevedo, I., Brown, R., Koomey, J., Masanet, E., Sartor, D., Herrlin, M., Lintner, W. (2016), *United States data center energy usage report*, LBNL-1005775. Berkeley, CA: Lawrence Berkeley National Laboratory. 65 pp. Available [viewed 2021-11-03] at: https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf.

[b-Shebabi, 2018] Shehabi, A., Smith, S.J.,, Masanet, E., Koomey, J. (2018). Data center growth in the United States: Decoupling the demand for services from electricity use. *Environ. Res. Lett*. **13**, 124030. Available [viewed 2021-11-03] at: https://iopscience.iop.org/article/10.1088/1748-9326/aaec9c/pdf.

[b-Song] Song, J., Ma, Z., Thomas, R., Yu, G. (2019). Energy efficiency optimization in big data processing platform by improving resources utilization. *Sustain. Comput. Inform. Syst.* **21**, pp. 80-89, https://doi.org/10.1016/j.suscom.2018.11.011.

[b-Strubell] Strubell, E., Ganesh, A., Andrew McCallum, A. (2019). Energy and policy considerations for deep learning in NLP, In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, pp. 3645–3650. doi: 10.18653/v1/P19-1355. Available [viewed 2021-11-08] at: https://aclanthology.org/P19-1355.pdf

[b-Synergy] Synergy Research Group (2020). *Hyperscale data center count reaches 541 in mid-2020; Another 176 in the pipeline*. Reno, NV: Synergy Research Group. Available [viewed 2021-11-10] at:: https://www.srgresearch.com/articles/hyperscale-data-center-count-reaches-541-mid-2020-another-176-pipeline

[b-Thibodeau] Thibodeau, P. (2017). Data centers decline as users turn to rented servers. *Computerworld*. Available [viewed 2021-11-03] at: https://www.computerworld.com/article/3188885/data-centers-decline-as-users-turn-to-rented-servers.html

[b-Thompson] Thompson, N.C., Greenewald, K., Lee, K., Manso, G.F. (2020). *The computational limits of deep learning*. Available [viewed 2021-12-28] at: https://cbmm.mit.edu/sites/default/files/documents/2020-07-10%20Thompson-Greenewald-Lee-Manso%20-%20Deep_Learning_Limitations%20-%20Neil%20Thompson.pdf

[b-Valcarenghi] Valcarenghi, L., Van, D.P., Raponi, P.G., Castoldi, P., Campelo, D.R., Wong, S.-W., Yen, S.-H., Kazovsky, L.G., Yamashita, S. (2012). Energy efficiency in passive optical networks: where, when, and how? *IEEE Network.* **26**(6), pp. 61–68. doi: 10.1109/MNET.2012.6375895.

[b-Veitch] Veitch, P., Broadbent, A., van Rossem, S., Sayadi, B., Natarianni, L., Al Jammal, B., Roullet, L., Mimidis, A., Ollora, E., Soler, J., Pinniterre, S., Paolino, M., Ramos, A., Du, X., Flouris, M., Mariani, L., Riganelli, O., Mobilio, M., Shatnawi, A., Orru, M., Zembra, M. (2018). Re-factored operational support systems for the next generation platform-as-a-service (NGPaaS). In: *2018 IEEE 5G World Forum* (5GWF), pp. 1-5. New York, NY: Institute of Electrical and Electronics Engineers. doi: 10.1109/5GWF.2018.8516995.

[b-Whitehead] Whitehead, B., Andrews, D., Shah, A., Maidment, G. (2014). Assessing the environmental impact of data centres – Part 1: Background, energy use and metrics. *Build. Environ.* **82**, pp. 151–159. Available [viewed 2021-11-04] at: https://reader.elsevier.com/reader/sd/pii/S036013231400273X?token=E0B9739489102C901FA6B9F991B720D906D0CA72045E4EF01D97E9C563244555975B2B7D9816A2698D491D2801FB1848&originRegion=eu-west-1&originCreation=20211104061434

[b-Zafrir] Zafrir, O., Boudoukh, G., Izsak, P., Wasserblat, M . (2019). *Q8BERT: Quantized 8bit BERT*. arXiv:1910.06188 [cs.CL].

[b-Zhang] Zhang, Y., Chowdhury, P., Tornatore, M., Mukherjee, B. (2010). Energy efficiency in telecom optical networks. *IEEE Communications Surveys Tutorials*, **12**, pp.441–458. doi: 10.1109/SURV.2011.062410.00034.