Electronic Devices & Networks Annex EDNA

Policies for Data Centre Energy Efficiency: Scope, Trends and Availability of Data

FEBRUARY 2023

Technology Collaboration Programme
by iea

Energy Efficient End-use Equipment
International Energy Agency

The Technology Collaboration Programme on Energy Efficient End-Use Equipment (4E TCP), has been supporting governments to co-ordinate effective energy efficiency policies since 2008.

Fifteen countries have joined together under the 4E TCP platform to exchange technical and policy information focused on increasing the production and trade in efficient end-use equipment. However, the 4E TCP is more than a forum for sharing information: it pools resources and expertise on a wide a range of projects designed to meet the policy needs of participating governments. Members of 4E find this an efficient use of scarce funds, which results in outcomes that are far more comprehensive and authoritative than can be achieved by individual jurisdictions.

The 4E TCP is established under the auspices of the International Energy Agency (IEA) as a functionally and legally autonomous body.

Current members of 4E TCP are: Australia, Austria, Canada, China, Denmark, the European Commission, France, Japan, Korea, Netherlands, New Zealand, Switzerland, Sweden, UK and USA.

Further information on the 4E TCP is available from: **www.iea-4e.org**

Electronic Devices & Networks Annex
EDNA

The EDNA Annex (Electronic Devices and Networks Annex) of the 4E TCP is focussed on a horizontal subset of energy using equipment and systems - those which are able to be connected via a communications network. The objective of EDNA is to provide technical analysis and policy guidance to members and other governments aimed at improving the energy efficiency of connected devices and the systems in which they operate.

EDNA is focussed on the energy consumption of network connected devices, on the increased energy consumption that results from devices becoming network connected, and on system energy efficiency: the optimal operation of systems of devices to save energy (aka intelligent efficiency) including providing other energy benefits such as demand response.

Further information on EDNA is available at: **iea-4e.org/edna**

This report was commissioned by the EDNA Annex of the 4E TCP. It was authored by Dr. D.H. Harryvan of Certios. The views, conclusions and recommendations are solely those of the authors and do not state or reflect those of EDNA, the 4E TCP or its member countries.

Views, findings and publications of EDNA and the 4E TCP do not necessarily represent the views or policies of the IEA Secretariat or its individual member countries.

The IEA, 4E, EDNA and the authors make no conclusions, endorsements or disendorsements in relation to any organisations or brands mentioned in this report.

POLICIES FOR DATA CENTRE ENERGY EFFICIENCY: SCOPE, TRENDS AND AVAILABILITY OF DATA

Created by Certios
For EDNA

VERSION | STATUS	1 | ready for publication
DISSEMINATION LEVEL	OPEN
AUTHORS (PARTNER)	D.H. Harryvan (Certios B.V.)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
1	Completed and reviewed	23-01-2023	Dr. D.H.Harryvan, Certios

EXECUTIVE SUMMARY

This report was commissioned by the EDNA Annex of the IEA 4E Technology Collaboration Programme. It proposes a definition and characterization of data centres, presents trends in their energy use, and analyses the availability of data which might be used as inputs to energy efficiency metrics.
The aim of this report is to assist policy makers in developing and implementing policies for various data centre types.

Many data centre standards and KPI's exist, but few are directly useful for characterizing data centres . Most are internal metrics that help a data centre or IT equipment owner to select and operate equipment.

This report concludes that only three distinct data centre types need be distinguished. These are based on who has responsibility for energy consumption:

	Optimization parameters		
Configuration	*Building supporting infrastructure*	*IT equipment*	*Software*
Enterprise DC	Data centre owner		
Co-hosting DC	Co-hosting provider		Co-hosting customer
Co-location DC	Co-location provider	Co-location customers	

Other distinguishing characteristics such as size and service level appear to be less relevant for the purpose of energy efficiency policies.

For the foreseeable future, the energy efficiency improvement of IT equipment, referred to as Koomey's law (Koomey, 2011) is expected to continue, although at a much slower rate than has occurred in past decades. The utilization of the IT equipment is expected to rise, due to the proliferation of cloud services. Increased utilization contributes significantly to higher energy efficiency.

For data centre cooling infrastructure, it is expected that air cooling (currently the preferred method of cooling IT equipment) will remain as such for the near future. Direct liquid cooling, which has many technical advantages, is expected to show significant growth, however its broad adoption is limited by a lack of standards, high investment costs and limited choices of liquid cooling compatible IT equipment.

Since IT equipment improvements are expected to lead to higher heat densities, there is a distinct possibility that recommended cooling temperatures in data centres will be lowered back to around 20 °C from the current recommendation of 27 °C. Lowering air temperature inside the data centre limits the use of free cooling technology and consequently results in lower cooling efficiencies. Separation of high density systems from the general server population, or targeted use of liquid cooling technology, could counteract this effect.

Overall, the current trend of increasing energy use by data centres is likely to continue. The expectation being that the growth in demand for computing resources will outpace any efficiency gains.

In order to address rising energy demand, the collection of relevant high-quality data is paramount. The study of data centre trends in this report revealed that there are many uncertainties, not only about future developments, but also about the status quo. There is significant variation in the estimated energy used by data centres, a condition that could be remedied by registering all data centres and collecting information on their operation.

This document further discusses data availability, considering several proposed metrics: two functional efficiency metrics, the IT Equipment Average Efficiency Index (ITAEI) and the Data Centre Functional Efficiency (DCFE), as well as a non-functional metric, the Data centre idle coefficient (DcIC) in combination with Power usage effectiveness (PUE).

This report found that reliable, publicly available information on the energy used by data centres is scarce. However such data is generally collected by data centre operators for internal use, and could be made available. Additional data, such as functional capacity, efficiency and utilization of specific functions of IT equipment are currently unavailable, and with the exception of CPU utilization statistics, are not generally collected. This makes obtaining such data very difficult.

Consequently, this report concludes that detailed functional metrics, such as the proposed ITAEI and DCFE - that are information rich but need detailed data - are impractical.

The DcIC, a metric that considers only non-functional energy use (energy lost on idle cycles in servers) is more practical since it relies on information that is currently monitored. The metric, while imperfect, can bring necessary insight in the operation of IT equipment inside the data centre (note that it currently only considers server energy). When combined with PUE, the DcIC, will quantify inefficiency, giving insight as to how energy can be saved and providing the ability to track improvements. Determination of the DciC is however not trivial, but possible.

All three of the above metrics have merit but need further development. In order to foster adoption of (in)efficiency metrics and promote data collection and availability, these developments should be done with close collaboration of the data centre industry.

This report also recommends that an effort be made to assemble a complete and continuously updated data set with the following basic information for data centres:

- Location
- Data centre type (Enterprise/co-location/co-hosting)
- energy use and PUE
- contact information

In addition, an effort should be made to acquire data on the operational (in)efficiency of complete data centres.

TABLE OF CONTENTS

1. INTRODUCTION

With the data centre (DC) industry growth comes an ever-increasing demand for energy. It is therefore important to ensure that the energy that is provided to the data centre is used in an efficient manner. In order to assess whether energy is used efficiently, it is first necessary to define what a data centre is, what its boundaries are and what types of data centres exist.

This work is part of the “Data centre Energy Efficiency policy workstream” started by the 4E Electronic Devices and Networks Annex (EDNA). 4E is an international platform for collaboration between governments, providing technical analysis and policy guidance to its members and other governments concerning energy using equipment and systems.

Within EDNA a workstream on energy efficiency of data centres has been established (EDNA, 2022).The goal of this workstream is to provide policy makers with information and evidence-based recommendations for policy measures to improve the energy efficiency of data centres, including the impact of these measures and suggestions for implementation.

1.1. Objective for the report

The objectives for this Activity 1 report “Definition of Scope, Trends and Data Availability and Quality” are to:

- Assess the various definitions and classifications for data centres, including but not limited to the uptime institute tiers, EN 50600 norms classification, the power usage effectiveness (PUE), the Data centre Energy Management (DCEM) from ETSI, type of data centres (colocation, enterprise, public, cloud, edge…). Propose a definition of scope and a classification of data centres, suitable for policy measures on energy.
- Identify trends in data centres, including the main components that constitute a data centre, related to energy efficiency. Where applicable, relate trends to the proposed classification.
- Assess the data availability (both in current practice and in future and including amongst others the measurability) for the metrics discussed in the EDNA project on Data Centre Metrics as well as for the proposed classification, including an assessment of the quality.

2. DATA CENTRE DEFINITION AND CATEGORIZATON

Since not all data centres are equal and cannot be treated as such, it is helpful to describe categories for data centres that are suitable to cover the entire market, yet sufficiently narrow so that possible policies can be effective for the targeted category.

To create such data centre categories, a literature study has been conducted of:

- Building codes
- Data centre standards
- Code of Conducts for data centres
- DC Guidelines
- Scientific literature on DC.

2.1. Data Centre definition

The first step is to define what is a "Data Centre" and to indicate what are the boundaries of the data centre wherein the potential policies should be proposed.

An often-quoted source for data centre related topics is the American Society of Heating Refrigerating and Air conditioning Engineers, technical committee 9.9 (ASHRAE TC9.9, 2022). TC 9.9 has approved the following definitions:

- *Data Centre*: A room or building, or portions thereof, with one or more *ITE enclosures* with a power input greater than 2 kW.
- *ITE Enclosure*: A rack, cabinet, or chassis that is designed to mount ITE.
- *Information Technology Equipment (ITE)*: Computers, data storage, servers and network / communication equipment.

The American ENERGY STAR glossary (ENERGY STAR, 2022) defines a data centre as:

- *Data centre* refers to buildings specifically designed and equipped to meet the needs of high-density computing equipment, such as server racks, used for data storage and processing. Typically, these facilities require dedicated uninterruptible power supplies and cooling systems. Data centre functions may include traditional enterprise services, on-demand enterprise services, high performance computing, internet facilities, and/or hosting facilities.
  Often data centres are free standing, mission critical computing centres. When a data centre is located within a larger building, it will usually have its own power and cooling systems and require a constant power supply of 75 kW or more. Data centre is intended for sophisticated computing and server functions; it should not be used to represent a server closet or computer training area.
- *Gross Floor Area* should include all space within the building(s) including raised floor computing space, server rack aisles, storage silos, control console areas, battery rooms, mechanical rooms for cooling equipment, administrative office areas, elevator shafts, stairways, break rooms and restrooms. When a data centre is located within a larger building, it includes only the spaces that are uniquely associated with the data centre in the gross floor area. For example, do not include spaces shared by the data centre and other tenants, such as break rooms or hallways.

These definitions, while very similar, still leave room for discussions. The discussions focus on boundary definitions. Boundaries exist on the outer perimeter of the "data centre", determining where the data centre starts, and inside the data centre, determining where the "data centre" ends.

While in general terms the data centre often denotes both the structure that supports the IT equipment and the IT equipment contained therein, the definitions above might suggest that the "data centre" is only the structure that supports the IT equipment and does not include the IT equipment itself.

Alternative definitions that incorporate the elements from the first two definitions and do include the IT equipment contained within the data centre are proposed in the Energy Efficiency Directive of the EU council (General Secretariat of the Council of the european union, 2022):

- '*Data centre'* means a structure, or group of structures used to house, connect and operate computer systems/servers and associated equipment for data storage, processing and/or distribution, as well as related activities.

The definition provided by ETSI documentation on Data Centre Energy Management (DCEM) (ETSI, 2014), which is also copied into the EN50600 technical report 99-1 "Recommended practices for energy management" (CLC/TC 215, 2021), is as follows:

- *Data centre:* 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.

2.2. Data Centre boundaries

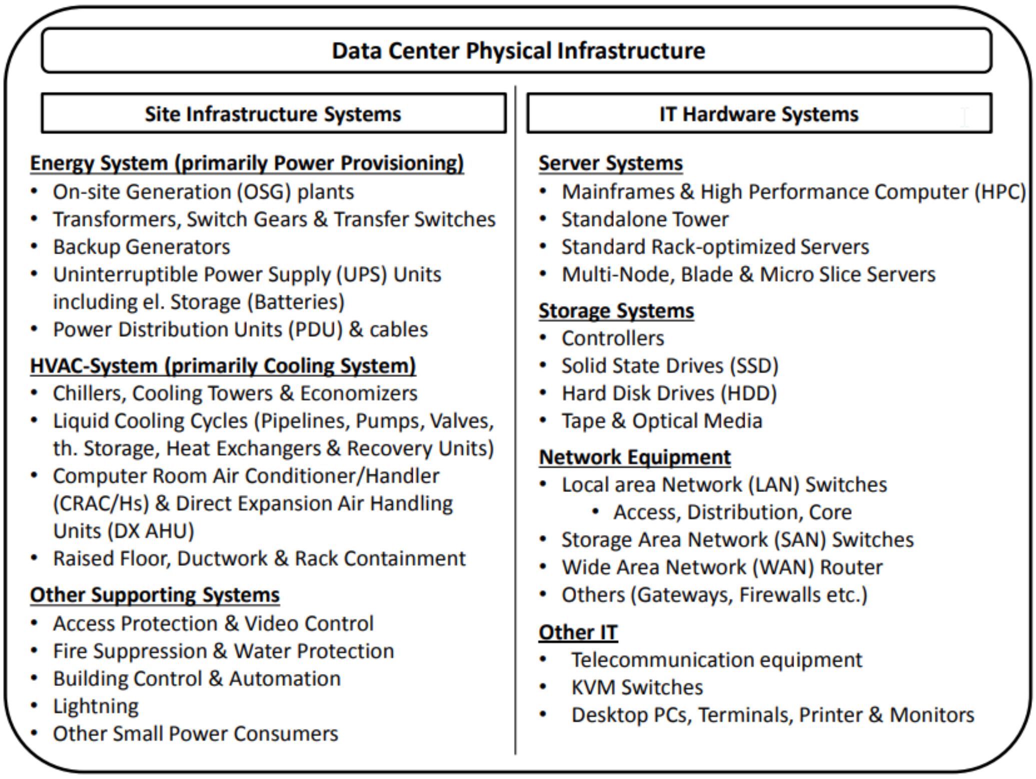
From the content of the ETSI and EN50600 documentation, which contain specific references to IT equipment, the intent of the ETSI definition concerning the data centre's boundaries becomes clear. Internally, there is no boundary defined. The "data centre" reaches all the way to the IT equipment contained within the data centre, including this IT equipment and the functions performed by the IT equipment. To avoid misinterpretation, we propose a slight modification to the ETSI definition (in **bold**)

- *Data centre:* structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of, **and including,** 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.

The lack of an internal boundary is of importance because certain measures such as those affecting environmental control, for example physical placement of the racks and maintaining separation between supply air and return (exhaust air) used for equipment cooling, will be the responsibility of the data centre operator. For measures such as choosing efficient IT equipment, workload management or software choices, different parties may be responsible.
As will be shown later, the addressees for measures can be the operator or the data centre customer (i.e., possibly but not necessarily different parties).

The ETSI definition does give a clear indication of an outside boundary of the data centre.
The data centre includes all the facilities and infrastructure for power distribution, environmental control, resilience and security as far as it is needed for the primary function of the data centre, namely the centralized accommodation and operation of the IT equipment.

These boundaries are also considered in other sources, e.g. **Figure 1** is taken from the paper "Data centre Green Performance Measurement: State of the Art and Open Research Challenges" (Schödwell, 2013).This paper describes an infrastructure-oriented approach to define the data centre and its boundaries.



Data Center Physical Infrastructure

Site Infrastructure Systems

Energy System (primarily Power Provisioning)
- On-site Generation (OSG) plants
- Transformers, Switch Gears & Transfer Switches
- Backup Generators
- Uninterruptible Power Supply (UPS) Units including el. Storage (Batteries)
- Power Distribution Units (PDU) & cables

HVAC-System (primarily Cooling System)
- Chillers, Cooling Towers & Economizers
- Liquid Cooling Cycles (Pipelines, Pumps, Valves, th. Storage, Heat Exchangers & Recovery Units)
- Computer Room Air Conditioner/Handler (CRAC/Hs) & Direct Expansion Air Handling Units (DX AHU)
- Raised Floor, Ductwork & Rack Containment

Other Supporting Systems
- Access Protection & Video Control
- Fire Suppression & Water Protection
- Building Control & Automation
- Lightning
- Other Small Power Consumers

IT Hardware Systems

Server Systems
- Mainframes & High Performance Computer (HPC)
- Standalone Tower
- Standard Rack-optimized Servers
- Multi-Node, Blade & Micro Slice Servers

Storage Systems
- Controllers
- Solid State Drives (SSD)
- Hard Disk Drives (HDD)
- Tape & Optical Media

Network Equipment
- Local area Network (LAN) Switches
  - Access, Distribution, Core
- Storage Area Network (SAN) Switches
- Wide Area Network (WAN) Router
- Others (Gateways, Firewalls etc.)

Other IT
- Telecommunication equipment
- KVM Switches
- Desktop PCs, Terminals, Printer & Monitors

Figure 1 : Data Centre physical infrastructure (Schödwell, 2013).

While this this is an attractive point of view for defining measures that address efficiency issues for specific machinery, it is not sufficient to address the interplay between the two domains or the innovations within each domain. The description also seems to exclude software as part of the data centre infrastructure, which as discussed earlier, should be included.

A different boundary is described in the ISO 30134 documentation describing Data centre key performance indicators (KPI) (ISO/IEC JTC 1/SC 39, 2016) (ISO/IEC JTC 1/SC 39, 2021). In these documents the data centre´s boundary is defined based on energy flows, as shown in **Figure 2**. The documents define data centre total energy; supporting the notion that a data centre is defined including the IT equipment. Further examination does show that the IT equipment is separated from the data centre´s infrastructure. According to these documents, IT energy should be measured separately and as close to the IT equipment as possible. This separation however does not represent an internal logical boundary as such, instead it supports the concept that in some cases, the parties responsible for the IT infrastructure and those responsible for the data centre infrastructure could be different.
The difference in responsibilities for certain parts of the data centre will be discussed in paragraph 2.3. under data centre categorization.

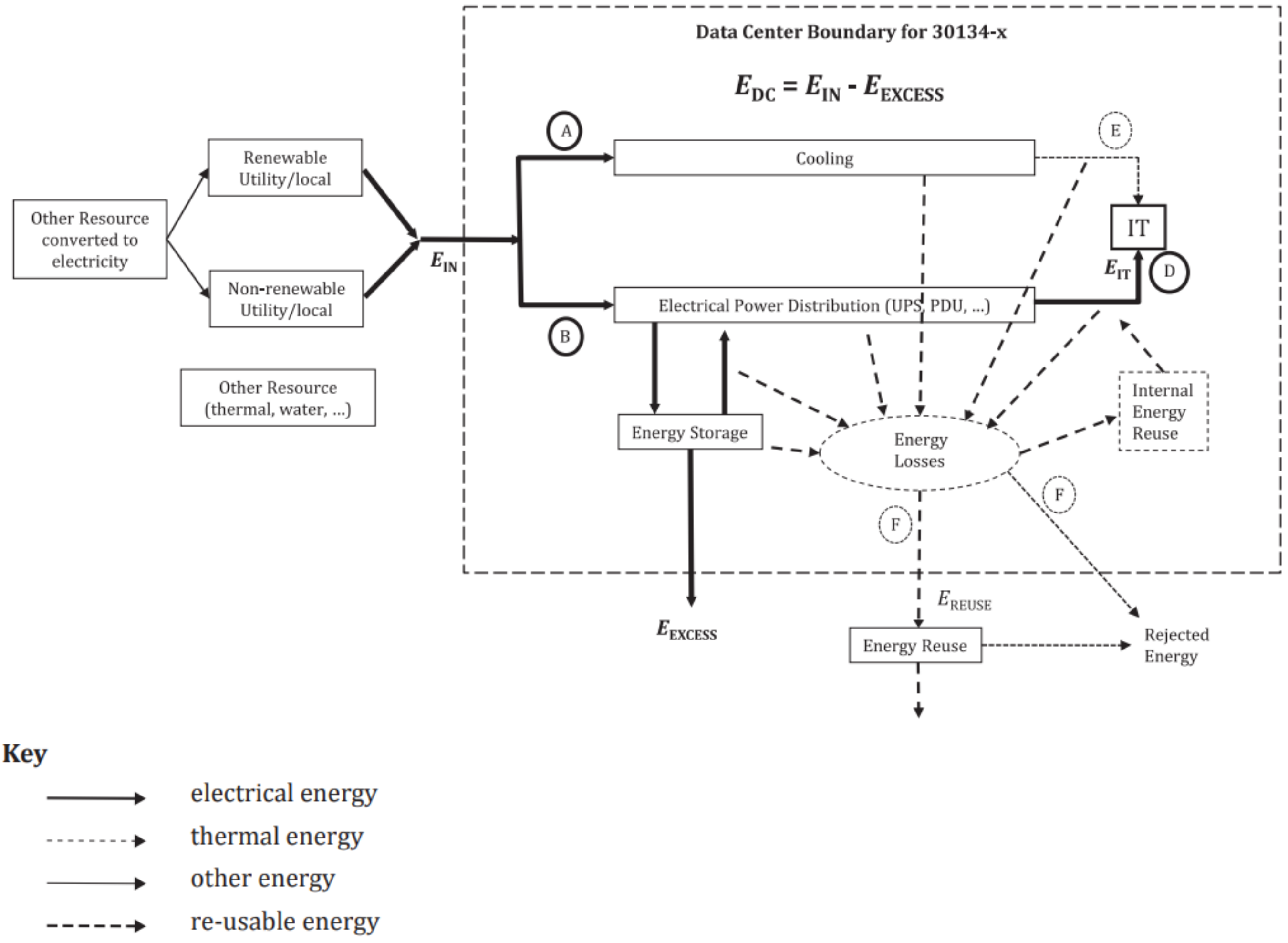


Figure 2 : Data centre´s outside boundary based on energy flows (ISO/IEC JTC 1/SC 39, 2021).

As shown in **Figure 2**, the DC boundaries as considered in ISO 30134-x are defined based on energy flows. All utility equipment not dedicated to the DC is outside of its boundaries.
The approach in the ISO documentation for setting the boundaries seemingly contradicts the ETSI and the paper by Schödwell, in the accounting of energy delivered in other forms than electrical energy. Its is important to realize that this difference is only the result of simplification needed for the determination of KPIs. The issues arising from this simplified view are directly visible in the calculation of the power usage effectiveness (PUE). The PUE (ISO/IEC JTC 1/SC 39, 2016) is a commonly used metric that represents a ratio between energy used by IT equipment in the data centre and all other energy use. The PUE is based on measuring electrical energy only. Modern climate control systems for DCs often rely on water evaporation to provide cooling, which replaces the electrical energy used in mechanical cooling equipment and therefor results in a lower PUE. Similarly, a DC can employ co-generation (combined heat and power) plants, use absorption chillers that covert heat into cooling, or use chilled water from other sources.

After reviewing the available documentation the amended ETSI definition seems to be the most suitable to adhere to, in the sense that it is the energy component that is most relevant for the current scope; and all energy and material flows such as electrical and thermal energy, water and fuel are also considered:

- *Data centre*: structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of, and including, 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.

The lack of a boundary between the data centre and the IT equipment, including the services provided by this equipment, contained within the data centre should not deter a policy maker from creating policies that primarily affect the IT equipment.
Measures exist that will improve the energy efficiency of the IT equipment itself and such measures can be enforced with policies. The distinction between data centre infrastructure and IT infrastructure only indicates that it is possible that the organisation/party responsible for the IT equipment is different than the organisation/party responsible for the data centre.

Similarly, setting the outer boundary at the data centre should not inhibit exploring policies beyond the data centre, e.g, for reusing waste heat from data centres. Heat reuse can still be an effective measure, but waste heat reuse does not directly contribute to the primary function of the data centre as described in the definition. The equipment needed for harvesting the heat (or increasing, if needed, the temperature) are outside of the boundary of the data centre.

2.3. Data centre characterization

Having established the existence of data centre boundaries this section discusses further the characterization of various types of data centres.

A starting point for characterization can be found in the technical report prepared by the European Committee for Electrotechnical Standardization CENELEC (CLC/TC 215, 2021), in which the following information is provided:

*"The **implementation** of data centres varies in terms of:*

a) purpose (enterprise, co-location, co-hosting, or network operator facilities);
b) security level;
c) physical size;
d) accommodation (mobile, temporary and permanent constructions).

***The needs** of data centres also vary in terms of*

e) availability of service;
f) the provision of security;
g) the objectives for energy efficiency.

These needs and objectives influence the design of data centres in terms of building construction, power distribution, environmental control, telecommunications cabling and physical security as well as the operation of the data centre. Effective management and operational information are important in order to monitor achievement of the defined needs and objectives."

Of these characterization aspects, e) availability c) size a) purpose (in order of importance), are commonly used in the data centre industry, to indicate differences between various data centres. Each of these aspects is discussed next.

2.3.1. Data centre availability of service

Within the industry, data centres are first and foremost distinguished based on the availability of service that they provide. Availability of service for the entire data centre is influenced by the design of the data centre site infrastructure systems for power and cooling as described in figure 1. Availability is generally

increased by adding additional power and cooling equipment which influences the overall efficiency of a data centre.
A number of standards for the availability design of the DC exist, such as:

- Uptime institute (UI) Tier classification system (Uptime Institute, 2022)
- ANSI/TIA "rated levels" (TIA, 2017)
- EN 50600 "classes" (CENELEC, 2019)

These standards reflect similar requirements for four availability classes, also referred to as "Tiers". These classes can be applied to any or all of three categories:

a) Power supply and distribution;
b) Environmental control;
c) Telecommunications cabling.

A Class 1 solution (single path):

- A single fault in a functional element can result in loss of functional capability; planned maintenance can require the load to be shut down.

A Class 2 solution (single path with redundancy):

- A single fault in a device shall not result in loss of functional capability of that path (via redundant devices);
- Routine planned maintenance of a redundant device shall not require the load to be shut down.

  NOTE Failure of the path can result in unplanned load shutdown and routine maintenance of non-redundant devices can require planned load shutdown.

A Class 3 solution (multiple paths providing a concurrent/repair operate solution):

- A fault of a functional element shall not result in loss of functional capability;
- Planned maintenance shall not require the load to be shut-down;
- For environmental control: although a failure of a path can result in unplanned load shutdown, maintenance routines shall not require planned load shutdown as the passive path serves to act as the concurrent maintenance enabler as well as reducing the recovery of service time (minimizing the mean downtime) after the failure of a path. All paths shall be designed to sustain the maximum load.

A Class 4 solution (fault tolerant solution except during maintenance):

- A fault of a functional element shall not result in loss of functional capability;
- Planned maintenance shall not require the load to be shut-down;
- For power supply and distribution: any single event impacting a functional element shall not result in load shut-down;
- For environmental control: a failure of one path shall not result in unplanned load shutdown All paths shall be designed to sustain the maximum load.

Since the focus of this document is on energy efficiency, this paragraph will discuss the impact of availability designs on the expected energy efficiency of a data centre.
in order to understand the influence of the data centre site infrastructure on efficiency, a short introduction on the most commonly used metric, the power usage effectiveness (PUE) is needed.

The PUE is defined as: $\mathrm{PUE} = \frac{\text{Total DC energy consumption}}{\text{Total IT hardware energy consumption}}$

For a correct determination of the PUE, energy is measured over a full year.
The total DC energy consumption can be written as the sum of the energy consumed by the IT hardware equipment contained in the data centre (total IT hardware energy) and the energy used or lost in all equipment needed to run the data centre (site infrastructure energy) as described in **figure 1**. (Schödwell, 2013)**.** Consequently, the PUE indicates how much energy the DC site infrastructure uses for each kWh that the IT hardware uses.
For example, a PUE of 1,5 indicates that if the IT equipment uses 1000 kWh in a year, the site infrastructure components add 500 kWh for a total energy consumption of 1500 kWh.

The categories that are mentioned in the class definitions are a direct match to the energy consuming elements that are part of the site infrastructure of a data centre. Different classes have different designs for crucial functions such as electrical power distribution and safeguarding and environmental control. Creating highly redundant infrastructure, up to the levels of Class 4, is complex and costly. However with regards to energy efficiency, the effect of redundancy on system load for the various components is what matters.

a. Power supply and distribution

The documents defining classes, dictate the design for the electrical distribution including all equipment for switching and safeguarding the supply of electricity to the IT equipment.
The class design dictates redundancy of distribution pathways and redundancy in backup generators and uninterruptible power supplies (UPS). The nature of the IT equipment is such that even short interruptions of the electrical supply can cause havoc. As a consequence, having backup generators is a necessary precaution but does not offer sufficient protection. In case of a power failure, backup generators will start within minutes, much to slow to prevent shutdown of the IT equipment. To bridge the gap between the power failure and start of backup generators, data centres deploy UPS systems. UPS systems include an energy storage, often batteries, and provide the ability to deliver power to the IT equipment within milliseconds after a primary supply failure.
While there are several different technological solutions for creating a UPS, all UPS systems introduce energy losses in the electrical path. The losses are a function of the amount of energy that passes through the UPS and are generally expressed as UPS efficiency:

$$\text{UPS efficiency} = \frac{Power\ output\ (kW)}{Power\ input\ (kW)}$$

The actual efficiency of UPS systems is load dependent. In general, highest efficiency is achieved for loads over 25% of the maximum achievable load, as shown in **Figure 3** (oura, 2016).

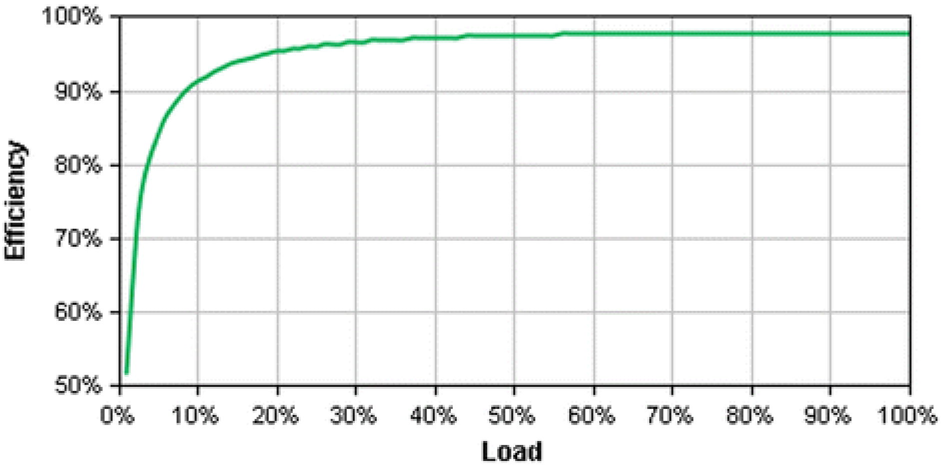


Figure 3: Typical UPS efficiency curve (oura, 2016).

UPS efficiency relates to the previously described classes through this load dependency. In theory, a Class 1 data centre will have a single UPS to safeguard the electrical supply to all IT systems. When the data centre is full, the UPS will have a 100% load.

In comparison, a fault tolerant Class 4 data centre will have a minimum of 2 active UPS systems. Each of these will carry at most 50% of the maximum load.

Since data centres are rarely filled to maximum capacity actual ups loads are often far below maximum. Class 4 data centres can therefor suffer from low UPS efficiency.

While this reasoning would indicate that data centre classes are relevant for characterizing data centres in terms of energy efficiency, technological advances have alleviated the issues of underloading UPS systems. Modern UPS systems are comprised of modules and are therefor scalable. The use of appropriately sized modules allows a data centre to scale UPS systems to match the IT load, reaching efficient, i..e. above 25%, loading of their UPS systems.

The second element on the class definition concerns the electrical distribution pathways

Electrical resistance loss in cabling is inversely proportional to the rated capacity of the cable. Simply put, thicker cables have lower losses. Electrical resistance loss is lowest in Class 4 DC, since all cabling is sized to handle full electrical load, while in normal operations the total load is spread evenly over multiple distribution paths. The maximum load in a normally operating Class 4 DC is below 50% of the rated capacity, therefore the electrical loss of cables is low.

b) Environmental control

Next to the electrical supply systems, the most crucial active equipment that a data centre uses to support the IT equipment within the data centre is associated with the environmental control.

All the energy that is supplied to IT equipment is ultimately converted to heat. This heat needs to be removed from the data centre in order to maintain the climatic conditions that are dictated by the technical specifications of the IT equipment.

In other words, all data centres need cooling. Since most IT equipment is cooled with air, data centres employ fans to supply cool air to, and extract warm air from, the IT equipment.

Additionally, depending on the type of cooling that a data centre uses, compressors and pumps are used to create cooled liquid and distribute this to where it is needed.

In modern data centres the use of compression based cooling, also known as “direct expansion (DX) cooling” is limited. Therefor, the electrical energy consumed in the environmental control infrastructure is mostly attributable to pumps and fans, driven by electrical motors.

There is a strong relationship between the speed (flow) and power of a fan or pump which is described by the “affinity laws”. The affinity laws simplify when speed is the only variable, according to the following equation (engineering toolbox, 2022):

$$P_1 / P_2 = (n_1 / n_2)^3$$

In which P_1 stands for the power needed to drive the fan or pump at speed n_1 and P_2 stands for the power needed to drive the fan or pump at speed n_2. For fans and pumps, the amount of air/liquid that is moved within a certain time is called “flow”, this flow is linearly proportional to the speed.
The equation is depicted graphically in **Figure 4**; the speed is directly proportional to the flow (blue line), while the required power is proportional to the third power of the flow (yellow line).
This means that the efficiency (displacement per unit of energy) of a fan or pump gets better when the flow decreases.

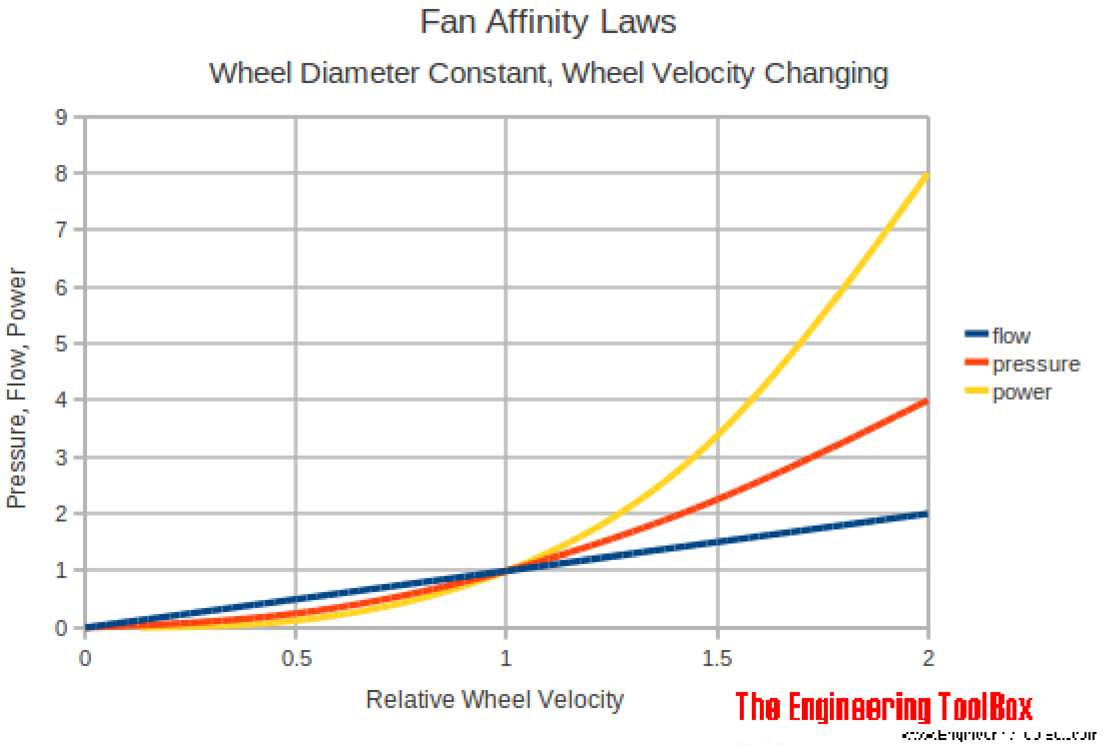


Figure 4 : Relation between motor speed and power draw in fans and pumps (engineering toolbox, 2022).

As with the electrical systems, Class 3 and 4 data centres have active redundant systems. Consequently pumps and fans will run well below their maximum capacity (flow).
In practice, most motors will have minimum speeds to ensure that pressure is maintained. Operation under severe underload can result in inefficiency (as with the previously discussed UPS systems). This underloading can simply be prevented by a scalable system design in which surplus capacity can be switched off.

c) IT configurations and software

Currently there are no standards to describe high available (HA) data centres created through the use of software techniques and IT equipment, like those created by a specific physical infrastructure (such as described by DC Classes before). Most of the high availability needs though can be fulfilled through a combination of IT configurations and data centre physical infrastructure design.
There are many possibilities for ensuring availability of services using virtual, self-healing IT configurations, that can span multiple data centre locations which do not need to be Class 3 or 4.

The impact of high availability on the energy efficiency of DCs can range from significant to almost negligible. This depends on the configuration used and the faults or even disasters that must be covered by such configuration.

For example, for banking applications, where account and transaction information are managed in multiple, geographically separated data centres, the traditional HA cluster configuration used supports the highest availability and covers failure of the entire data centre. These HA clusters consist of multiple hardware nodes, distributed over multiple sites, of which only one is actively processing data. The other nodes are also powered, but their function is copying the data from the active node and monitoring whether the active node is still functioning.

High availability can also be provided by multiple active nodes in so called "N+1 configurations". In these cases, a limited amount of headroom is maintained to cover the failures of a single (virtual) node or component, and the additional energy demands are rather small.

A review of business strategy and the need and method for attaining high availability is a part of the best practices described in the EU Code of Conduct (EU-JRC, 2022).

In the characterization of data centres attaining HA through IT configuration will probably influence the need for high Class data centres, but would not require itself a separate, additional classification of the data centres.

In summary, the data centre classes (tiers) are used by the data centre industry to distinguish between various types of data centres. Nevertheless these Classes do not imply a direct relationship to the energy efficiency of the data centre.

Instead, the technical discussion above points towards the fill factor, that is to say the actually used capacity as fraction of the maximum capacity, and the adaptability of the data centre to changes in IT load as determining influences on data centre efficiency. A prime example of this fact is shown by the Swisscom AG Class 4 data centre in Bern. This data centre, realized in 2015, employing a modular build strategy reports a PUE value of 1,22 (Heslin, K, 2015).

2.3.2. Data centre size

Data centre size is most commonly expressed in capacity parameters, either as available floor space for IT equipment or as available electrical power for IT equipment. Recent developments in high density computing place more emphasis on the electrical power than on floor space.

In terms of electrical power, the capacity of data centres ranges from kW to over 100 MW per location. There are no clear standards for data centres of different size or capacity, but literature reports the following terms:

- Server room/Server closet: these names generally indicate small locations that are not separated from the facilities (power and cooling) of the buildings in which these are located. As such they fall outside our definition of data centre as "a structure dedicated to accommodate IT equipment".
- "Edge" data centres: this is a relatively new term that often means that the data centre is small, but the term really denotes the location of the data centre in the data network. As the name suggests, edge data centres are located near network boundaries, close to the source and use of data streams.
- "Hyperscale" data centres: the name suggests large sites, but again a formal definition is not available. The market research company Synergy Research Group (Synergy group, 2022) uses the term hyperscale to refer to companies rather than to individual data centres. According to Synergy research, hyperscale operators run hundreds of thousands of servers in their data centres. The largest cloud services vendors have millions of servers (Leopold, 2017). The market

research company Gartner states that "*A hyperscale vendor is a very large-scale cloud provider with global reach and impact*" (Karamouzis F., 2022), so the number of servers is not considered in this case.

With respect to energy efficiency, the size of the data centre seems marginally relevant. A 2017 study of data centres completed for the EU CoC for data centres, showed that there was no significant variation in PUE between data centres of different sizes (Avgerinou, Bertoldi, & Castellazzi, 2017).
However, size shall not be entirely ignored. While the measures that ensure efficient operation work just as well in smaller data centres as in large data centres, the economies of scale of the large data centres allows for investments that will not be economically feasible for smaller data centres. A distinction between small or medium sized and (very) large data centres might be needed. Large data centres could operate under more stringent KPIs and require additional efficiency measures. The reason being the impact of a large data centres on its surroundings (impacts such as Land, Energy and Water usage as well as Waste heat management). The economies of scale of the large data centres allows for investments that would not be economically feasible for smaller data centres.
One example of a measure targeted at large data centres could be waste heat reuse. Connecting many small data centres to a district heating system might be too costly, where connecting a single large data centre would make economic and environmental sense. At the time of writing of this document, Germany is considering a new energy efficiency act that mandates the re-use of waste heat. In this act the minimum size for data centres that are subject to the obligations in the law is proposed at 100 kW. (eversheds-sutherland, 2022)

There is however no literature support or market consensus on what defines small/medium/large data centres and combined with the observations from Averinou, Bertoldi & Castellazzi (Avgerinou, Bertoldi, & Castellazzi, 2017), we conclude that setting categories based on size is not a practical, robust approach concerning data centre´s energy efficiency because size will not be the most relevant factor.

2.3.3. Data centre purpose **(enterprise, co-location, co-hosting, or network operator facilities)**

The data centre purpose, as defined by the European Norm (EN) 50600 (CENELEC, 2019) refers to the ownership and responsibility on the different components that constitute an operational data centre and the IT equipment which it contains. *Purpose* in this case does not refer to the work done by the IT equipment inside the data centre.

The following definitions are given in the norm documentation:

- **Co-hosting data centre:** is a data centre in which multiple customers *are provided with access* to network(s), servers, and storage equipment to operate their own services and applications. The information technology equipment and the supporting infrastructure in the building (such as power distribution and environmental control) are provided as a service by the co-hosting data centre operator/owner(the co-hosting provider), and (most of) the software is operated by (multiple) co-hosting customers (lessees).
- **Co-location data centre:** is a data centre in which multiple *customers locate their own network(s), servers and storage equipment*. In this case, the supporting infrastructure of the building is provided as a service by the co-location data centre operator/owner (the co-location provider), (multiple) co-location customers provide their own hardware (servers, storage and network equipment) and software.
- **Enterprise data centre:** is a data centre operated by an enterprise which has the sole purpose of delivering and managing the services for its employees and customers. In this case, the building, hardware and software all have the same owner/operator; the data centre owner.

- **Network operator data centre:** is a data centre for the delivery and management of broadband services to the operators' customers. The network operator data centre is a subset of the enterprise data centre.

There are compelling reasons for proposing a characterization based on the ownership/operation of a data centre, e.g., considering the following aspects:

- The operation of the building infrastructure
- The operation of the hardware
- The operation of the software.

Targeting the policies to parties according to their responsibilities and ability to comply to a policy target would be possible. This is the case in the European Code of Conduct for Energy Efficiency in Data Centre (EU CoC) (EU-JRC, 2022).

The EU CoC, lists best practices and recommendation for energy efficiency measures, mapped onto functional elements and responsible parties. According to the EU CoC, an IT service is built upon several functional elements, for which different parties may bear responsibility, as shown in **Table 1.**

Table 1 : EU CoC components for providing an IT service.

Area	Description
Physical building	The building including security, location and maintenance.
Mechanical and electrical plant	The selection, installation, configuration, maintenance and management of the mechanical and electrical plant.
Data floor	The installation, configuration, maintenance and management of the main data floor where IT equipment is installed. This includes the (raised) floor), positioning of air-conditioners and air-handlers units and basic layout of cabling systems (under floor or overhead).
Cabinets	The installation, configuration, maintenance and management of the cabinets into which rack-mount IT equipment is installed.
IT equipment	The selection, installation, configuration, maintenance and management of the physical IT equipment.
Operating System / Virtualization	The selection, installation, configuration, maintenance and management of the Operating System and virtualization (both client and hypervisor) software installed on the IT equipment. This includes monitoring clients, hardware management agents, etc.
Software	The selection, installation, configuration, maintenance and management of the application software installed on the IT equipment.
Business practices	The determination and communication of the business requirements for the data centre including the importance of systems, reliability availability and maintainability specifications and data management processes.

The first column in Table 1 shows the various components that are ultimately needed for providing an IT service, and the description gives information about where possible measures (policies) could be considered.

The EU CoC also defines 5 types of parties as included in **Table 2**. These parties are held responsible for the corresponding components of the functional stack previously shown in **Table 1**.

Table 2 : Type of party according to the EU-CoC (ref?)

Type of Party	Description
Operator	Operates the entire data centre from the physical building through to the consumption of the IT services delivered.
Co-location provider	Operates the data centre for the primary purpose of selling space, power and cooling capacity to customers, who will install and manage their own IT hardware and services.
Co-location customer	Owns and manages IT equipment in a co-location data centre, in which they purchase managed space, power and cooling capacity.
Managed service provider (MSP)	Owns and manages the data centre space, power, cooling, IT equipment and some level of software for the purpose of delivering IT services to customers. This could also include traditional IT outsourcing.
Managed service provider in co-location space	A managed service provider which purchases space, power and cooling in a data centre, to provide services to third parties.

While apparently differing from the EN50600 classifications, one can map the Code of Conduct parties to definitions included in the EN50600, as shown in **Table 3**.

Table 3: Mapping parties according to EU Coc and EN50600.

Parties acc. to EU CoC	Acc. to EN 50600
Operator	Enterprise data centre
Co-location provider	Co-location data centre
Co-location customer	-
Managed service provider (MSP)	Co-hosting data centre
Managed service provider in colocation space	-

Combining both standards gives us a working characterization of data centres based on the parties responsible for implementing certain energy efficiency measures.

Table 4 : proposed data centre characterization

Category	**Optimization parameters**		
	Building supporting infrastructure	*IT equipment*	*Software*
Enterprise DC	Data centre owner		
Co-hosting DC	Co-hosting provider		Co-hosting customer
Co-location DC	Co-location provider	Co-location customers	

Since in first instance, the focus of this document is on supporting infrastructure and IT hardware energy efficiency, the situation for co-location data centres is different. Although co-location providers may be able to set some requirements regarding IT hardware, e.g., to ensure that this can be safely installed and operated, they may not know the energy efficiency and operational conditions of that hardware. Even more, the co-location customers may not want that the co-location provider knows this information. This means that for this type of DC both, the co-location provider and the co-location customers, could be the addressees for (policy) measures. An alternative could be that regardless of the legal requirements concerning the ownership of the building and/or hardware, the legal persons in control over the building and/or hardware are the addressees of any kind of information or performance (energy efficiency) requirements.

2.4. Summary of the data centre characterization

As described in Chapter 2, the amended ETSI definition of a data centre seems the most suitable for the EDNA DC workstream, because it is based on the energy aspects of DCs. All energy and material flows such as electrical, thermal, water and fuel are included in this definition.

Data centre: structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of, and including, 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.

A classification including three types of data centres is proposed, based on the ownership/operation of a data centre and considering the following components:

- Operation of the building infrastructure
- Operation of the IT equipment
- Operation of the software

The responsibility on these components can be in hands of different parties. Three common cases are:

- *Enterprise data centre*: the building, hardware and software all have the same owner/operator, the data centre owner.
- *Co-hosting data centre*: the building and the IT hardware are provided by the data centre operator/owner (the co-hosting provider), (most of) the software is operated by (multiple) co-hosting customers (lessees).
- *Co-location data centre*: the building is provided by the data centre operator/owner (the co-location provider), (multiple) co-location customers provide their own hardware (servers, storage and network equipment) and software.

3. TRENDS OF ENERGY USE IN DATA CENTRES

3.1 Overall trends

The comprehensive report “International Review of Energy Efficiency in Data Centres” discusses the projections for the energy use of the global data centre market (Brocklehurst, 2021).
The report mentions several uncertain factors that can have an large effect in terms of the projected energy usage by he DC industry, namely:

- The growth in data traffic
- The adoption of 5G and the consequent development of edge data centres
- The development and adoption of energy efficient technology and business practices

Brocklehurst collected a number of projections from various sources and shows that these result in widely differing predictions for the year 2030. The predicted energy use varied between 250 TWh up to 3000 TWh annually in 2030.

A 2019 IEA Commentary “Data centres and energy – from global headlines to local headaches? “ , explored the global and local energy implications of data centres and shows that there is a global trend on moving capacity from smaller “traditional”enterprise and co-location data centres to large-scale or “hyperscale” data centres (IEA, 2022). The commentary also presents energy use data for DCs, as shown in **Figure 5**.

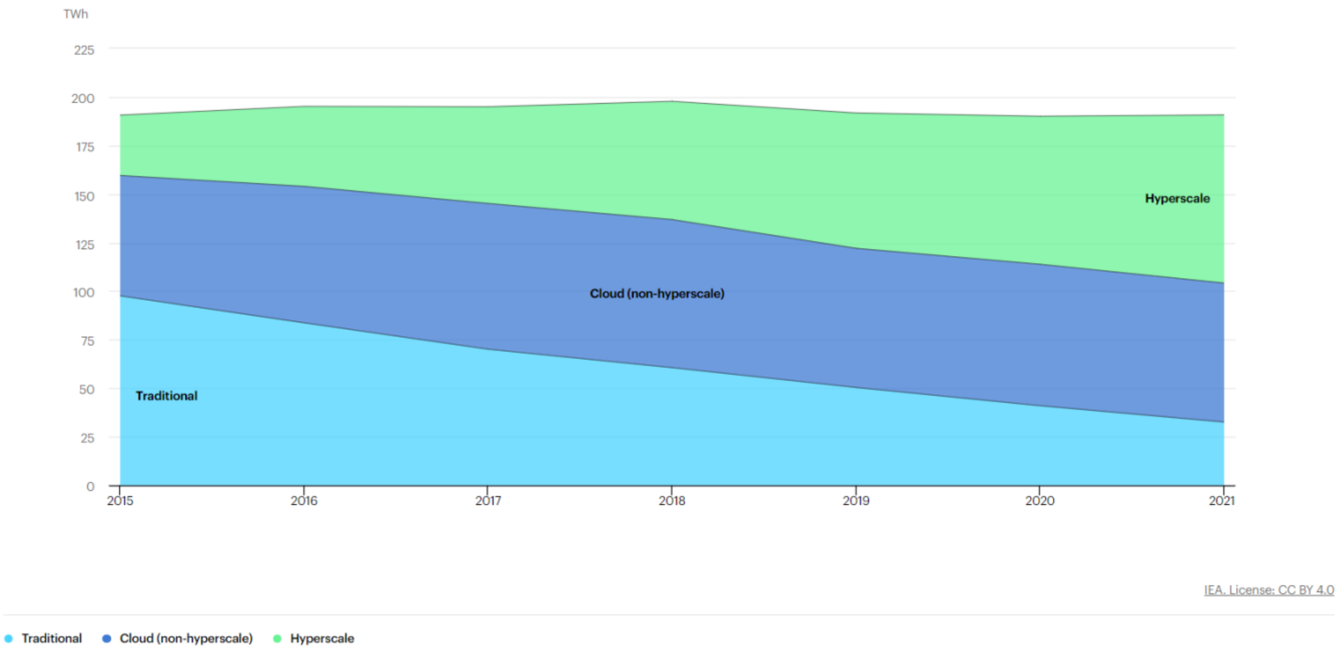


Figure 5 : Global data centre energy demand by data centre type (TWh) by type (IEA, 2022)

IEA, Global data centre energy demand by data centre type, IEA, Paris https://www.iea.org/data-and-statistics/charts/global-data-centre-energy-demand-by-data-centre-type, IEA. License: CC BY 4.0

Unfortunately, reliable data about worldwide energy usage by data centres is limited, and data uncertainty remains high. The recent report Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market also illustrates this data uncertainty. The estimates for the worldwide data centre energy use in 2018 ranged from 200 TWh to over 600 TWh annually, with a CAGR of 10% (Montevecchi, 2020).

The IEA 2019 Commentary mentions cooling efficiency as a reason for the lack of growth in energy demand in light of an increased demand for IT services. In contrast, a study by 451 research, (Bizo, 2019) attributes the larger contribution to efficiency to the fact that the cloud business model of a shared and monetized infrastructure drives server utilization well above what is possible for enterprises. Bizzo asserts that most of the energy savings come specifically from two factors: more energy efficient servers and much higher server utilization. This contrasts with the frequent references in literature and promotional materials that point to the low energy use of the infrastructure in such data centres, resulting in low values of PUE.

In our view, both data centre site infrastructure systems as well as IT contribute to energy efficiency improvements, and as such, developments in both areas will be discussed in the following paragraphs. Despite efficiency gains, ongoing digitalization and especially the increasing availability of cloud services will lead to a significant growth in data centre capacities. This expected growth is so strong that it will more than off-set the significant efficiency gains achieved at all levels (hardware, software, and data centre infrastructure). The strong growth in light of increased efficiency is often referred to as Jevons paradox (Bauer & Papp, 2009), and as a result, the total energy consumption of data centres is still expected to rise (Montevecchi, 2020).

The share of edge data centres will also increase significantly in the future. According to a report prepared for the European Commission, edge data centres are expected to account for 12% of the energy consumption of data centres in EU28 (Montevecchi, 2020).

The expectations for data centre energy use are very dependent on the scenarios used for doing the forecasts. Demand growth and efficiency gains will continue to be important drivers impacting energy use. In the worst-case scenarios, energy use by data centres can more then double over the coming 10 years, in the best-case scenarios, efficiency gains will outpace demand growth, and energy use will drop significantly. The technological developments indicate that scenarios where energy use will continue to rise are more likely for the near future.

3.2 IT Equipment trends

To assess the IT equipment trends, data on the distribution of energy use associated with the four major elements inside the data centre is available, for example from the 2019 IEA Commentary (IEA, 2022), as shown in **Table 5**.

Table 5 : Energy use of functional elements of the data centre

Elements inside the DC	Estimated energy use 2021 (TWh)
DC site infrastructure	59
Servers	109
Network	4
Storage	19

Given the fact that the DC site infrastructure energy use is dominated by the cooling, which in turn is directly coupled with the energy use and heat losses of the IT equipment, one could draw the conclusion that servers are responsible for over 80% of the total energy use of a data centre. It is for this reason that the computational performance of servers is of paramount importance to the energy efficiency of the entire data centre. Two factors determine this performance: the equipment utilization and the hardware efficiency.

Server utilization

Modern servers show a strong relationship between electrical power draw and utilization. One source of data that illustrates this relationship is the SPECpower Benchmark, which reports measurements of server performance and power draw at various utilization levels. The results have been published since 2008 (Standard Performance Evaluation Corporation, 2022), and for illustration, latest result published during the writing of this report has been selected.

The test result for this system, ASUSTeK Computer Inc. Model: RS700A-E11-RS4U is shown in **figure 6**:

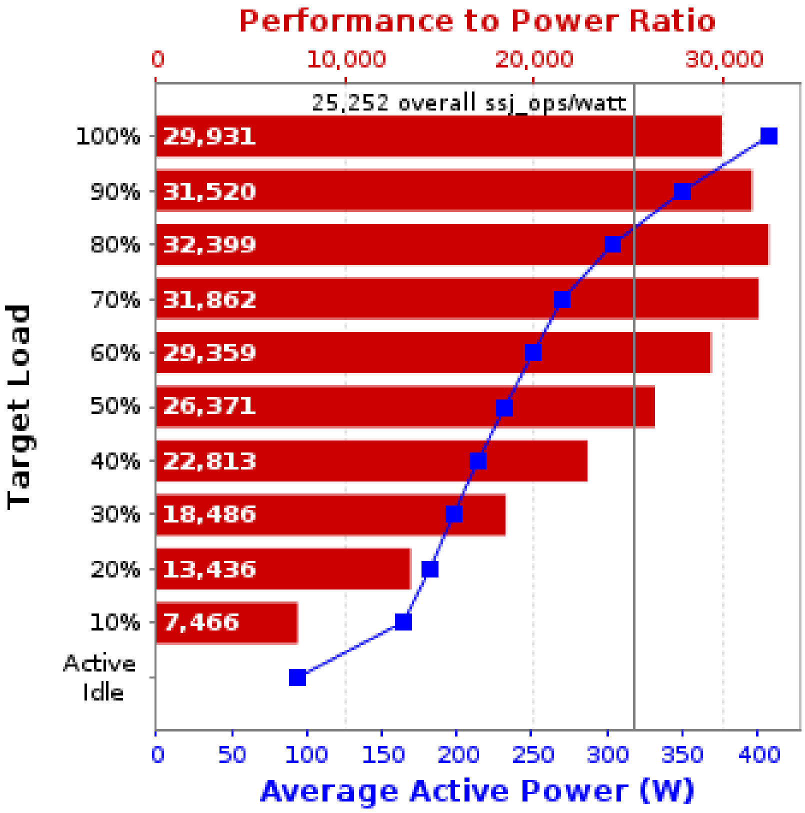


Figure 6 : SPECpower results for a 2-socket server 2022. (Standard Performance Evaluation Corporation, 2022)

In Figure 6, the blue squares show the electrical power draw at the utilization levels 0% to 100% in 10% intervals. More importantly, the red bars show the benchmark performance per Watt, which can be interpreted as computational efficiency.

It becomes clear that under low utilization (below 20%), the computational efficiency is 3 to 4 times lower than the maximal achievable computational efficiency that the same server can reach when optimally utilized (i.e., 80% load). Not only work output, but also efficiency increases with utilization. The result shown is typical for servers produced over the last decade. Load (utilization) is a key factor in efficiency, which explains why cloud services are delivered with much higher energy efficiency on average.

Taking the server system above as an example, this system is equipped with two CPUs, each consisting of 64 cores, each core supporting 2 compute treads. To fully load such a system, 256 compute treads must be active at any given point in time. Such an enormous demand for computer resources can only be achieved by virtualization, with a large number of active users. Even in such circumstances the average utilization rarely reaches levels above 50%.

Cloud services implies the sharing of resources between various cloud users. Virtualization software ensures the logical and secure separation between different users while enabling the use of a single server by multiple users. Because companies that offer clous services serve many users, it is possible to arrange the workload in such a manner that high average workloads are achieved on all hardware elements. This situation is hard to achieve by any single company deploying its own DC infrastructure.

Server efficiency

The SPECpower benchmark also illustrates the enormous efficiency advances made by IT hardware manufacturers. **Figure 7** below was published in February 2008, and like the results shown in **Figure 6,** these also correspond to the single server with 2 CPUs.

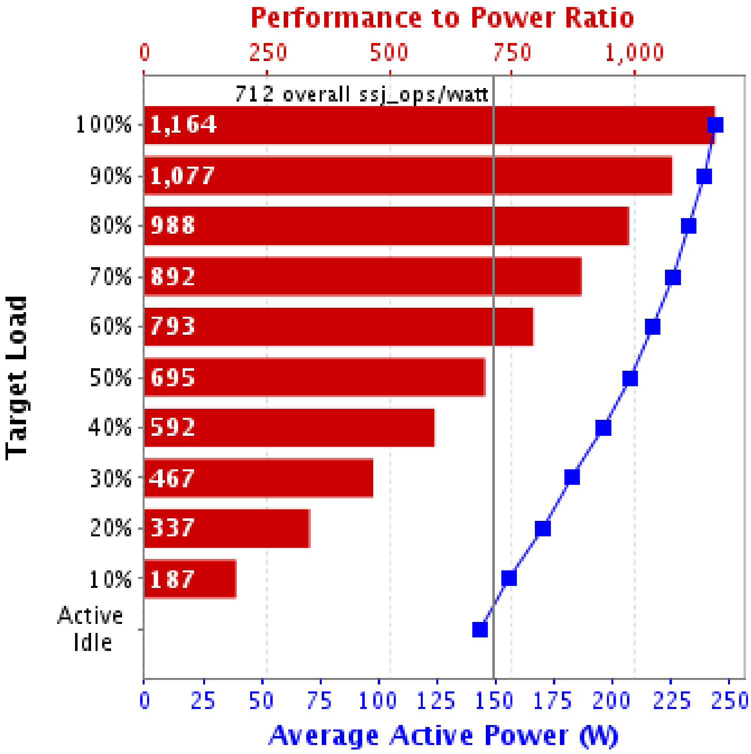


Figure 7 : SPECpower results for a 2-socket server 2008. (Standard Performance Evaluation Corporation, 2022)

From these results we can calculate that the average performance per Watt increased 35 times in a time span of 14 years. The performance per Watt under low utilization (10% and 20% load) has increased 40 times. This additional performance effect is mostly due to a lowering of the power draw for the “active idle” mode in modern equipment.

The main driver for improving the computing capacity to energy consumption ratio of servers over the past decades has been the continuous improvement and miniaturization of transistors that make up the CPU and other active components. In early years, the number of transistors on a chip doubled at regular

intervals. This was known as Moore's Law, at later date, these improvements were coupled to electrical power draw, and became known as Koomeys law (Koomey, 2011). Whether the fundamental physical conditions will allow for Moore's Law to continue, i.e., seeing further efficiency gains, is a subject of controversy (Montevecchi, 2020).
The improvements in computational efficiency are slowing down (Shalf, 2020). Transistors already reached an atomic scale (3 nm) (ASML, 2022) and further miniaturization will yield limited scaling, but this does however not signal an end of (other) system developments.
In May 2016, the International Roadmap for Devices and Systems (IRDS™) was initiated under IEEE sponsorship (IRDS, 2022).The IRDS publishes a roadmap of electronic developments, the latest publication is from 2021. The roadmap shows that transistor scaling will still contribute to efficiency gains. Additional gains will be achieved using other methods such as 3D packaging, placing logic/memory on top of each other rather than next to each other and the use of, Application Specific Integrated Circuits (ASICs), Graphics processing units (GPUs) and Field Programable Gate Arrays (FPGAs). In spite of these continued developments, the IRDS expects that the demand for IT services continues on an exponential growth path, which will result in a strong growth in the number of installed devices, and a strong growth in data centre capacity.

Truly new technology is being developed which will open a new era of computing power and efficiency. This technology is known as quantum computing. Commercially available quantum computing is not expected before 2030. These quantum computing systems will need cryogenic temperatures for cooling, so it is likely that quantum computing will emerge only in large scale, centralized data centres.

Virtualization trends

The need for improved utilization of IT equipment will continue in the IT industry. The initial response to the development of multi-core, multi-treaded volume servers was the adoption of virtualization technology, previously only available in mainframe and Unix based servers. Software “Hypervisors” allowed the building of virtual machines (VM), multiple instances of operating systems could now be run on a single physical server. The development of cloud computing with thousands of virtual machines deployed, sparked the trend of low overhead virtualization.
At the time of writing of this report, the use of “containerization” is already widespread as a virtualization technology. Containers share a single underlying operating system (OS), which in itself can be a virtual machine, diminishing the overhead of running many versions of the same OS. For a user, a container serves the same purpose as a virtual machine, a virtualized portion of a physical server is created to run an application, logically separated and secured from other users.
By virtue of running fewer instances of an OS, the containers are much more efficient than VMs, meaning that more application workload can be supported by the same IT hardware, using the same or even less energy.

The containerization is not the last step in overhead reduction. Technology is being developed that will reduce the overhead of the applications. This technology is confusingly called “serverless computing”. The term is confusing because obviously, a server is still needed to perform a computational function, what is indicated by the term is the notion that software developers do not need to consider OS or hardware restriction in the development and deployment cycle. “Serverless computing” is based on the concept of micro-services. The micro-service is a functional element of an application that is started and stopped separately. One can imagine that for example writing data to a hard disk is a function that is not frequently needed. As a micro-service, this functionality would only be started when needed, and stopped after use. A user functionality can be created by combining micro-services. A server would then only run the active components of an application (Johann Schleier-Smith, 2021).

The virtualization technology has not only touched the computational elements in the data centre, the development and use of network and storage virtualization has also taken flight, and adoption is growing. With more lightweight virtual environments tunning on a single physical server, more and more network communication happens within a server.
For this purpose, a virtualized network infrastructure is created, called “software defined network”, including switches, routers, firewalls etc; not limited by the physical limitation of a single server. Network management software handles the configuration of these networks, spanning many servers and including traditional network equipment to create a single coherent structure. This allows communication between selected senders and receivers wherever these may reside within the entirely virtualized environment, thus separating the networks of different users that are using the shared virtualized infrastructure.
As with networks, storage has also developed into a software defined structure. These structures will use both, server-based storage as well as specialised (storage) equipment, to store and safeguard data belonging to different users.

The continued virtualization of the entire IT infrastructure causes functional blurring. Where traditionally, functions such as data manipulation, data storage and data transport was done by specific hardware, servers now not only handle computational workload, but will also function as vital parts of the network and storage infrastructure. This functional blurring makes it harder to collect data needed for developing specific, functional metrics on energy efficiency.

In the coming decade IT systems will continue to improve in capacity as well as in efficiency. These improvements will however come at a slower rate compared with previous decades. These newer systems will most likely have higher energy demands, and more importantly, will create more heat in a smaller space than current systems. This increased heat density will have a negative impact on data centre cooling efficiency.
Over the last decades, improvement in capacity and efficiency of IT equipment, together with cooling efficiency improvements, have helped the IT industry to fulfil the growing demand for IT services and keep a limited increase in total energy use. However, if demand remains strong as expected, this future growth can only be met by a significant increase in the number of data centres, and an associated increase in energy use. The report “Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market” (Montevecchi, 2020) confirms this expectation for the European Union. The historical and expected energy consumption as proportion of the total electrical energy demand is shown in **figure 8.**

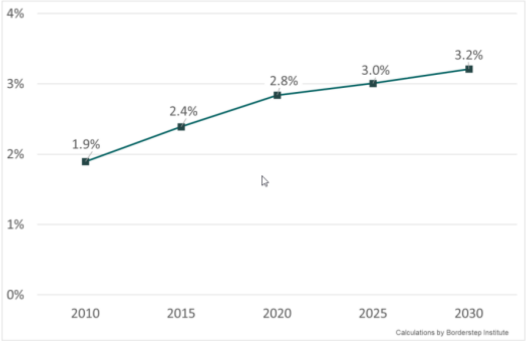


Figure 8 : Energy consumption of data centres between 2010 and 2030, trend scenario, as a proportion of EU28 final electrical energy demand. (Montevecchi, 2020).

Developments beyond this decade cannot be predicted with accuracy, quantum computing could bring a technological shift, but the commercial application of this technology is simply too far into the future for the scope of this EDNA workstream.

3.3 Data Centre cooling trends

The primary force that drives changes in data centre cooling technologies and operation are the IT equipment development directions. Developments such as 3D packaging and a drive to increase operational frequencies influence the need for cooling. Both these developments will cause an increase in heat density in the computational core of a server.
Other influences on data centre cooling are energy cost, internal company, environmental policies and external drivers such as environmental laws and policies enforced by governments.

Air cooling
At this point in time, the primary method for cooling electronic equipment is still forcing cooled air through the systems. As heat density increases so does the volume and ultimately the temperature of the air needed to provide this cooling.
The American Society for Heating Refrigeration and Air conditioning Engineers (ASHRAE TC9.9, 2022) is worth mentioning here because of its importance in defining the temperature and humidity ranges in data centre environments. The technical committee (TC) 9.9 is the most important group for data centre professionals within ASHRAE.
There are a number of useful publications by TC9.9, including “the Thermal Guidelines for Data Processing Environments 5th ed”. The complete documents are accessible to ASHRAE members, for reference the most commonly referenced psychrometric chart, showing temperature ranges for data centre operations is shown in the figure below.

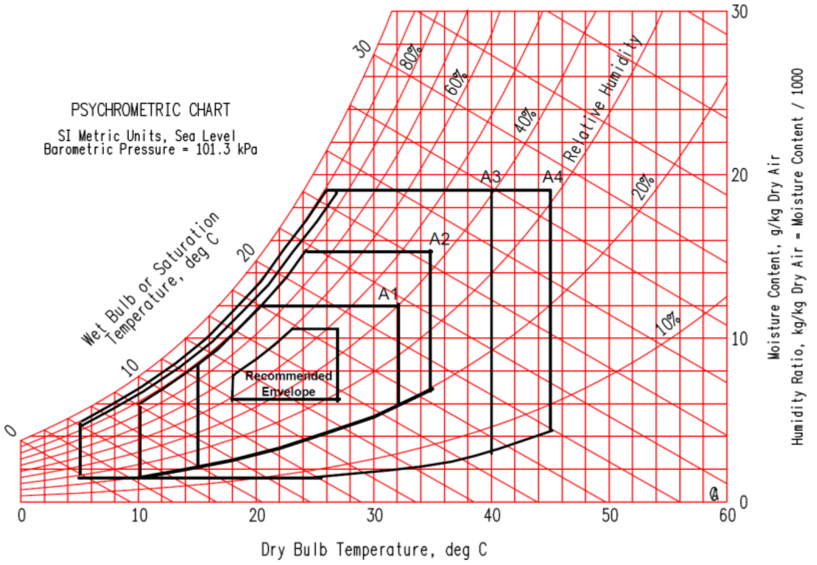


Figure 9 : ASHRAE A1-A5 (ASHRAE TC9.9, 2022)

The importance of this chart cannot be understated, IT equipment manufacturers rate their equipment in these classes and consequently, temperatures maintained in the data centres are related to these classes.
Most IT equipment (servers) comply to class A2, resulting in a maximum recommended cooling air temperature of 27 °C, and an allowable maximum of 35 °C. This in turn determines the free cooling range for a data centre and through this asserts a huge influence on cooling efficiency.
As discussed in the previous paragraph however, the trend in IT systems development will require a more stringent cooling regime. The ASHRAE seems to be aware of these developments and introduced a new “high density compute” class of systems H1. As shown in **figure 10**, for H1 systems, the maximum allowable cooling air temperature is 25 °C and recommended maximum is only 22 °C.

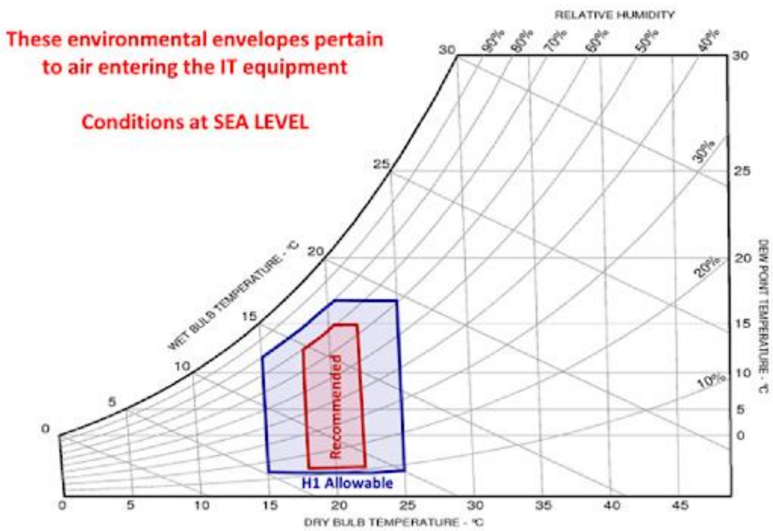


Figure 10 : ASHRAE H1 envelope (Bizo, New ASHRAE guidelines challenge efficiency drive, 2021)

When H1 IT equipment is installed in data centres, this will require the data centre to cool to 22 ^{o}C or below instead of the currently most accepted 27 ^{o}C or below.
Free cooling is the most important source for cooling efficiency to date. With free cooling, a data centre uses outside air to directly or indirectly cool the IT equipment. The method can only be used with favourable external conditions, having to cool below 22 ^{o}C severely limits this free cooling potential and therefor lowers the average cooling efficiency of the data centre.

Liquid cooling

The cumulation of increased heat density, the need for efficiency and in some cases the requirement for reusing data centre waste heat has caused a resurgence of direct liquid cooling technologies. Multiple variations of liquid cooling exist:

a) direct to chip
b) single phase immersion
c) two phase immersion

A complete discussion on the pro’s and cons of each of these technologies is beyond the scope of this document, but the market for liquid cooling shows a strong growth and while not totally replacing air cooling technology, direct liquid cooling will play a significant role in data centre cooling in the coming years.

In general direct liquid cooling has large advantages over air cooling.

1) Cooing Efficiency; Cooling either an entire system (immersion), or hotspots (CPU/GPU) within a server, with liquid is much more efficient compared with air. Heat transfer is faster, allowing more heat to be removed with much smaller heatsinks. This heat transfer efficiency allows for:
   - Higher IT equipment density resulting in more tightly packed IT systems and consequently less data centre space.
   - Higher heat output of hot components. This advantage is commonly exploited in the area of crypto mining where CPU overclocking produces more computational power but also more heat. CPU overheating is avoided by using a liquid cooled system.
   - Support for 3D packaging. In 3D packaging, logic components such as CPU ad memory are stacked, which means that heat generated in lower layers must travel through active higher layers before it can be removed. This setup requires the top surface of the package to be cooler then currently achievable with air cooling. Direct liquid cooling results in a

lower surface temperature and is therefor supportive for the developments in 3D packaging.

For future developments, liquid cooling facilitates the IT equipment developments but vice versa, the equipment developments also facilitate liquid cooling. One issue with liquid cooling is the presence of mechanical components such as hard disk drives (HDD). The liquid can interfere with the mechanics of spinning disks, these disks are gradually being replaced with solid state drives (SSD), a persistent memory-based solution without moving parts. These SSD's are currently still more expensive per GB storage than HDD, but offer more IO performance and are therefor favoured as internal storage solution in servers.

2) Temperature of waste heat; The enhanced heat transfer and heat capacity of liquids compared with air has additional benefits. One major advantage is that the temperature difference between the heat producer and cooling agent need not be so high. This temperature gradient is a driving force for heat transfer and is the reason that in air cooling, the air temperature must be below 27 ^{o}C in order to keep a CPU temperature within 70-100 ^{o}C.
The temperature differences with a liquid cooling medium are much smaller. Coolant intake temperature can be as high as 40 ^{o}C and output temperatures of 60 ^{o}C are easily attainable. These temperature ranges allow;
    - Free cooling year-round. With a simple dry cooler, 40 ^{o}C can be reached with free cooling only in almost all climatic zones, the year round.
    - Simple coupling with district heating (DH) for heat reuse. Since the heat carrying medium is liquid, transport over longer distances is less complex than with air as a heat carrier. The relatively high temperature makes direct use or coupling with a 4th generation district heating system simple and effective. The DH return is sufficiently cool for direct use in the data centre cooling (see **Figure 11**).

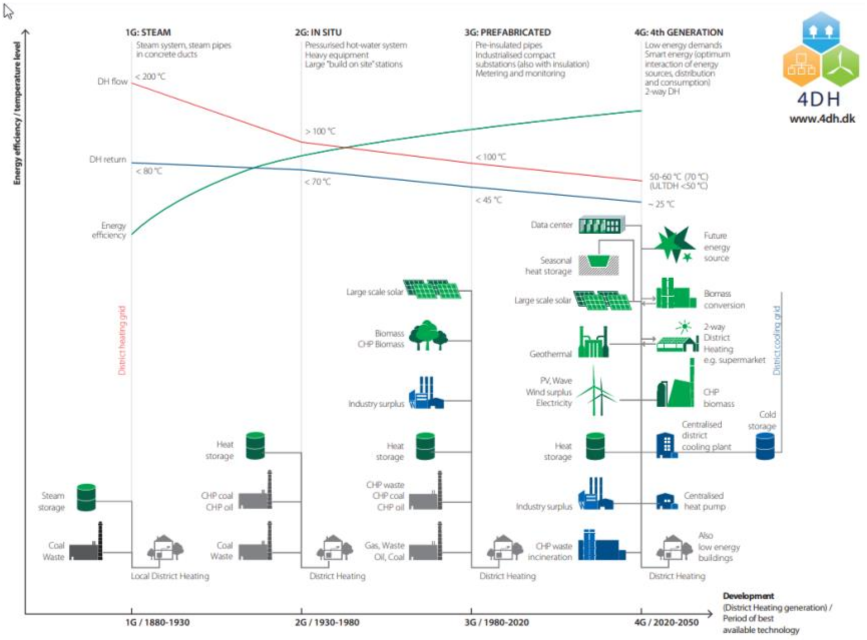


Figure 11 : Evolution of district heating systems (Thorsen, 2018)

3) IT equipment efficiency. The creation of cold air for use in air cooling of a data centre is not the only energy using component. In order to force air over the heat exchangers, IT equipment has their own fans installed. The energy use of the fans is counted as "IT energy" and therefor lowers the computational efficiency of a server when running at high speeds.

> In liquid immersion cooling, these fans are removed from the servers lowering the energy use of the server by 20-30% and thus increasing computational efficiency of the immersed server.

However, even with all of these advantages, liquid cooling is unlikely to dominate the data centre cooling market in the foreseeable future. Lack of standardization, high capital costs and limited availability of IT equipment that is prepared for liquid cooling will result in a slow adoption.

Conclusion

Air cooling will continue to dominate in data centre cooling, with direct liquid cooling showing significant growth. It is a likely scenario that set point temperature inside the data centre will be lowered in order to accommodate new IT technology. This set point lowering will result in less efficient cooling because the window for free cooling will be smaller.

In a favourable scenario, these lower setpoints (ASHRAE class H1) will be confined to separated areas within the data centre, the rest of the data centre will maintain the ASHRAE class A2 limits that are most common today. In such a scenario, the effect of lower cooling efficiency will be limited.

Looking beyond 2030, a highly specialized form of cooling might be needed to support quantum computers. Quantum computing requires cryogenic temperatures, specifically, the computer must be cooled with liquid Helium at approximately -270 ^{o}C. Creating and maintaining these temperatures has very little overlap with today’s data centre cooling technologies. It can be expected that specialized data centres will be created for this type of computing.

4. DATA AVAILABILITY

Policy creation with the aim of improving data centre efficiency is wholly dependant on the definition and determination of efficiency, which in turn requires data about the data centres at which the policy is aimed. Such data might not be easily accessible, the following paragraphs will discuss data availability in line with the data centre category definition discussed in the current document and data availability connected to selected metrics published in “Metrics for Data Centre Efficiency” a report prepared for IEA 4E EDNA by Viegand Maagøe in 2022. In this report, two metrics are proposed;

- the IT Equipment Average Efficiency Index (ITAEI)
- The Data Centre Functional Efficiency (DCFE)

The metric ITAEI would be dedicated to IT equipment and it evaluates the functional efficiency of the DC as the average used capacity divided by the average IT used power. In contrast, the metric DCFE evaluates the total work delivered and the total energy consumption of the DC along a reporting period. The different metrics are meant to respond to the needs from different policy options: for example, a policy option that requires information from DC operators will require different information, metrics and monitoring period than a policy option aimed at IT equipment owners or manufacturers.

A different approach to the creation of a metric is proposed In another report published by IEA 4E EDNA, “The Idle Coefficients” (D.H.Harryvan, 2021). In this report two alternative metrics are defined:

- The Server Idle Coefficient (SIC)
- Data Centre Idle Coefficient (DcIC)

The metric SIC is dedicated to servers and quantifies the percentage of energy used by a server that is spent on idle cycles. Idle cycles produce no result and occupy a server when there is no application workload. The DcIC elevates the SIC to the complete data centre, resulting in a percentage of the energy use of the IT equipment that is used for idle.

Optimization of the DcIC involves both equipment renewal, selecting equipment with a low idle power draw, but more importantly through avoiding idleness by workload placement optimization. As such, the DcIC addresses operation of the IT equipment by owners as well as design of the equipment by manufacturers.

4.1 Data related to data centre categories

Seemingly trivial and basic information, such as where data centres are located and the amount of data centres within one's borders, is not always available. Not all data centres make themselves known to a general public or the authorities. Registries common to most countries that hold information on commercial companies such as owner, industry code, VAT numbers and the like, offer some, but not total insight in the data centre market within a country. Data centres can be a part of a whole range of commercial activities and as such can be registered under virtually any industry code.

Easiest to identify and locate are the Co-Location data centres. Co-Location data centres need to attract customers and will commonly advertise name, location and contact information. Also, since Co-Location is the primary commercial activity, these data centres might well be traceable in local registries (tax and or legal).

The Co-Hosting data centres are harder to locate. Co-Hosting providers advertise their services, but not necessarily the location(s) from where these services are provided. A prime example where this is the case is Amazon Web Services Elastic Cloud, EC2 (AWS, 2022). Here a customer can procure virtual hardware resources and install his/her own software. There are over 20 regions that the customer can choose disconnecting the physical hardware location completely from the customers location.
In the case of large players, such as AWS, Google, Microsoft, Meta and others, the physical data centre locations are well known, but this openness is not guaranteed for many smaller providers that may or may not own a physical site.

Enterprise data centres are by nature of the fact that they only service a single company not inclined to advertise their presence. Their presence maybe inferred from the commercial activity of a company, financial services companies in general have large IT needs and commonly have their own data centres, but this inference does not equal hard data and cannot be relied upon.

To complicate matters further, although arguably size does not factor in either the data centre definition nor the data centre characterization, size does influence the availability of data on a particular data centre.
Simply put, small data centres are less easily found. An example of this can be found with cable network providers. These providers that provide internet and television services to the home have numerous locations that fall within the definition of data centre. Centralized locations, generally large, but many more small locations where the actual connection to the individual households originate. These locations are of course known to the cable company but these locations are not public knowledge.
It is to be expected (see “trends in data centres”) that with the roll out of 5G networks, a huge growth in such edge data centres will occur. These edge locations may be owned by telecom companies, making their locations more easily found, but just as likely have different owners that might be harder to track.

The availability of trivial data such as location and contact information but also energy use is limited for public authorities, (Brocklehurst, 2021) but the data is technically available. All of the DC locations are connected to an energy grid and data networks. Energy bills must be paid and equipment installed and serviced. For this, ownership, location, energy use, available capacity etc. is all information that is recorded and kept. Where the ownership of the site infrastructure systems differs from its user(s), customers are billed for the use of the data centre and as such their information is also recorded.

The European Union is working on resolving some of these issues with data availability. In the newest proposal for the Energy Efficiency Directive, (General Secretariat of the Council of the european union, 2022) the following addition is proposed:
Member States should collect and publish data, which is relevant for the energy performance and water footprint of data centres. Member States should collect and publish data only about data centres with a significant footprint, for which appropriate design or efficiency interventions, for new or existing installations respectively, can result in a considerable reduction of the energy and water consumption or in the reuse of waste heat in nearby facilities and heat networks. A data centre sustainability indicator can be established on the basis of that data collected
In addition:
Article 11a Data centres
Member States shall require, by 15 March 2024 and every year thereafter, owners and operators of every data centre in their territory with a significant energy consumption to make publicly available the information set out in Annex, which Member States shall subsequently report to the Commission

While many details will need to be worked out, it is apparent that the EU recognises the lack of publicly available relevant data on the data centre industry and is taking actions to remedy this situation by imposing an obligation on data centres to publish information.
If the proposal is accepted by the Member States and the European Parliament, the detailing might result in an addition to our current data centre classification: "*significant footprint*" implies that size might become a factor after all, but may also only be used to indicate a lower bound such as suggested earlier.

Conclusion: The information needed to identify, locate and categorize data centres according to the proposal is present, but not readily available for authorities. The most solid data resides within the combination of

- Utility providers, due to the often high capacity connection to the electrical grid.
- Network providers, due to the large number and high-capacity network connections to a data centre.

Unless such data is made available to the authorities, identifying data centres relies on data made public by the data centre itself. As such we conclude that the current data is most probably incomplete and of questionable quality.

4.2 Data related to metrics

Publicly available datasets on energy performance for any jurisdiction are limited. It is recognised that this does not support initiatives which are trying to increase sustainability. Typically, commercially owned or operated data centres do not publish energy performance data as they are commercially sensitive: for most, energy costs are their largest single operating cost. However, no evidence was found of any government *currently* requiring energy reporting by private sector data centres, even if that data were kept private. Mandatory reporting on energy use and relevant metrics by data centres is however considered by several Member States and the EU as a whole. The proposed German energy efficiency act and the French tertiary building act both require reporting. Reporting will also be mandatory if the new energy efficiency directive (EED) of the EU is adopted (General Secretariat of the Council of the european union, 2022).

Publicly available information includes:

- Statistics Netherlands has published data on the electricity supplied to data centres for 2017-2019 based on data provided by network operators.

- The French Agency for Ecological Transition (ADEME, 2022) has collected data in 2020 on the energy use and PUE of existing data centres within France. Furthermore, in accordance with legislation on obligatory energy reporting and improvements in energy efficiency, ADEME developed and launched the OPERAT platform (OPERAT, 2022). Data centres are classified as tertiary buildings in France as therefor fall under the law for mandatory reporting and improvements.
- In the US in 2018 the Energy Information Administration undertook a pilot for adding data centres as a separate building type to their periodic Commercial Buildings Energy Consumption Surveys (Energy Information Administration 2021) but found that obtaining data centre estimates was likely not feasible with current methods.

Further focused searches failed to uncover data on public (central or local Government) energy use. Some initiatives collect these data, for example the EU Code of Conduct and the US DCOI, but only rarely or partially publish the results. The data that have been found are limited and are generally reported in terms of PUE (Brocklehurst, 2021).

In the second half of 2022, EDNA is expected to have published the document “Metrics for Data Centre Efficiency”. This document not only examines existing metrics and frameworks for data centre efficiency, but also proposes metrics for use by policy makers, which are based on data measurements and collection at the individual DCs, useable as metric for development of DC energy efficiency at a national, regional or global level based on existing statistics.

4.2.1 IT Equipment Average Efficiency Index

The proposed IT Equipment Average Efficiency Index (ITAEI) includes used capacity for delivered useful work as an average over a measurement period. This metric would enable the evaluation of the IT functions delivered at different loads per average power consumed by the IT equipment. Therefore, it would not be a metric to monitor the operation of the DC, but to assess the efficiency of the IT equipment in terms of function delivered per power consumed by the IT equipment.

The three most common functions found in literature are processing, storage and networking. These three functions cannot be summed directly, they require to be normalised into an equivalent unit, and to this purpose, normalization factors are defined as the inverses of the power consumption efficiency of a Best Available Technology (BAT) server, BAT storage equipment and BAT network equipment. The ITAEI proposed equation would be as follows (av: average):

$$\mathrm{ITAEI} = \frac{\mathrm{N1x(servers\ av\ used\ capacity)\ +\ N2x(storage\ av\ used\ capacity)\ +\ N3x(network\ av\ used\ capacity)}}{\mathrm{av\ used\ power\ of\ IT\ devices}}$$

With:

- N1 = normalisation factor for servers = rated power of BAT server / rated capacity of BAT server
- N2 = normalisation factor for storage = rated power of BAT storage / rated capacity of BAT storage
- N3 = normalisation factor for network = rated power of BAT network / rated capacity of BAT network

The normalization factors build up a dimensionless index and since they reflect the BAT reference values, they serve as benchmarks with which to compare the used capacity and power. The value of the ITAEI ranges from 0 to 1, where 1 is the ideal value.

The definition of the normalisation factors requires technical analysis of a representative sample of equipment. The BAT reference of a data centre would consist of the set of three values for server, storage and network that represent the most appropriate equipment.

For servers such normalization factors have already been defined, the SPECpower benchmark mentioned in paragraph 3.2 is one example. However the SPECpower benchmark does not define the BAT, rather it allows manufacturers to demonstrate the performance per Watt of the servers they produce.

The use of “Best Available Technology” as a reference is both a strength and weakness of the ITAEI. As a strength, by defining a BAT reference the actual energy use of the IT equipment is linked to an ideal version of the equipment. This reference would evolve over time, consequently, an IT infrastructure comprised of aging IT equipment would score lower then an IT infrastructure that is continuously modernized.

The weakness of a BAT reference however is that, what constitutes “Best Available Technology” and who determines this is not determined. The technology with which to compare could differ from data centre to data centre, this could mean that it is the data centre that needs to define its own reference BAT. Additionally, while the use of a BAT reference encourages rapid replacement of equipment with more energy efficient equipment, such rapid replacement might not be desirable. Other, as yet unresolved discussions, point to more than just operational energy use of IT equipment. The production of IT equipment requires energy and several critical raw materials that cannot be reclaimed through recycling. These considerations indicate that long time use of IT equipment is preferable because of the reduction in raw material usage (Whitehead, 2015).

ITAEI is a metric based on an average of the results of testing used capacity and power over a sampling window, i.e., a representative period of time, hence it does not measure total work delivered with reference to the total energy consumption in that period of time. Testing is proposed to follow the recommendations from the Green Grid. The testing time would be between one to ten minutes aggregated into an average of two hours. Sampling windows should include peak bursts within each four-hour window per day during a 30-day interval.

4.2.2 Data Centre Functional Efficiency

The Data Centre Functional Efficiency (DCFE) measures the total functions delivered by a data centre and the total energy consumption over a reporting period. Contrary to the ETAEI, this metric focusses on the operation of the data centres, and includes all the energy consumed, not only IT equipment. Therefore, it is to be a tool for operators to monitor the total functional efficiency of the data centres, and for policy options aimed at collecting information by means of reporting provisions. The formula of DCFE would be similar to ITAEI, incorporating the total functions delivered and the total energy consumption over a period of time:

$$\mathrm{DCFE} = \frac{\mathrm{N1x(servers\ work\ delivered)\ +\ N2x(cumulative\ storage\ used)\ +\ N3x(cumulutive\ network\ used)}}{\mathrm{Total\ DC\ energy\ consumption}}$$

With: N1, N2 and N3 the same values as for ITAEI, in terms of rated capacity per power consumed by the BAT IT equipment, as it is a simple way to create a dimensionless index retaining the advantage of the BAT reference.

The DCFE would completely replace the PUE as a metric. The efficiency of the DC site infrastructure systems (mostly for cooling and power) that is indicated by the PUE is contained within the denominator: Total DC energy consumption. When a data centre would invest in more efficient cooling, this would result in a lower total DC energy consumption and automatically in a higher DCFE.

DCFE would need to measure total work delivered with reference to the total energy consumption in that period of time. Measuring the work delivered would entail testing in the same way as ITAEI, but instead of average, the measurement would be cumulative, first over the testing time, and then over the reporting period, which would be the same as the energy reporting period. The energy consumption needs to be measure along a year to take into account the effect of seasonal temperature variations.

4.2.3 Data Centre Idle Coefficient

Where the ITAEI and DCFE are efficiency metrics tied to the IT functionality, SIC and DCiC are metrics for inefficiency tied to energy waste. The DcIC is therefor similar in nature to the Power Usage Effectiveness (PUE). The PUE is currently the most commonly used metric to quantify the performance of the data centre site infrastructure systems. The PUE can be written as:

$$\mathrm{PUE} = \frac{\text{Total lT hardware energy consumption} + \text{Total DC site infrastructure energy consumption}}{\text{Total lT hardware energy consumption}}$$

The ideal value for PUE is 1, indicating that energy used by site infrastructure systems, that does not contribute to the productivity of the data centre, should be as low as possible and ideally zero. As discussed in the document “Metrics for Data Centre Efficiency”, the PUE is not an ideal metric. PUE only addresses the proportion of energy that is used by the site infrastructure and completely ignores any (in)efficiency of the IT equipment. This drawback of the PUE is well known, but the metric is very easy to obtain and does convey at least part of the information on the data centre efficiency. Because of this, PUE is used by almost all data centres to track and improve site infrastructure energy efficiency.

In order to overcome the drawback of the PUE, considerable effort has been expended to produce a functional metric that includes the IT efficiency. Many of the resulting metrics are discussed in the document “Metrics for Data Centre Efficiency”. Since all of the earlier metrics were deemed inadequate, the authors proposed the afore mentioned ITAEI and DCFE.

The DcIC was developed along the line of thought of the PUE development. Rather than trying to quantify useful work by the IT environment, the DcIC indicates the part of the energy use within IT equipment that indisputably does not contribute to the productivity of the IT equipment. This non-productive energy is used during idle cycles of a CPU, resulting in the following definition of the DciC:

$$\boldsymbol{DcIC} = \frac{\sum \boldsymbol{Server\ Energy\ (Idle)}}{\sum \boldsymbol{Server\ Energy\ (total)}}\ .100\%$$

The ideal value of the DcIC is 0%, indicating that idle energy use, that does not contribute to productivity, should be as low as possible and ideally zero. Since determining the DcIC involves the determination of the non-productive energy, it creates the opportunity to consolidate the DcIC and PUE in a single metric. (D.H.Harryvan, 2021) suggests an idleness corrected PUE (ICPUE) which could take the form of

$$\mathrm{ICPUE} = \frac{\text{Total lT hardware energy consumption} + \text{Total DC site infrastructure energy consumption}}{\text{Total lT hardware energy consumption}-\text{Total idle energy}}$$

Like PUE, ICPUE is a dimensionless ratio that is based on energy measurements. The ideal value would be 1.

Like all other metrics, the DcIC is not perfect. While the DcIC does not need definition of a best available technology (BAT) baseline, it also does not benefit from it. Equipment replacement only contributes to better DcIC when the power draw under idle diminishes. The metric is most sensitive for operational improvements, avoiding idle cycles altogether.

4.2.4 Data availability and quality for ITAEI, DCFE and DcIC

In the previous sections, metrics were proposed, which require all or part of the following data from a data centre:

- Used capacities of servers, storage and network
- Used power of servers, storage and network
- Used energy of the entire data centre
- Work delivered by servers, cumulative storage and network used along a reporting period.

In addition, to these measurements, when using the ITAEI and DCFE, for each of these categories agreement must be reached as to what is defined as “Best Available Technology” and the units in which the normalization factors are expressed.

When analysing data availability and quality, the first step is to split the data into two categories:

- Electrical data
- IT usage data

Electrical data

Focussing on the electrical data first: The document “Metrics for Data Centre Efficiency” (EDNA, 2022) includes a schema for where in the power distribution measurement points should be available in order to collect the required electrical data. According to the document, IT power should be collected at the tertiary distribution level, which is at the socket level, where the IT equipment is connected to the power grid (see Figure 12).

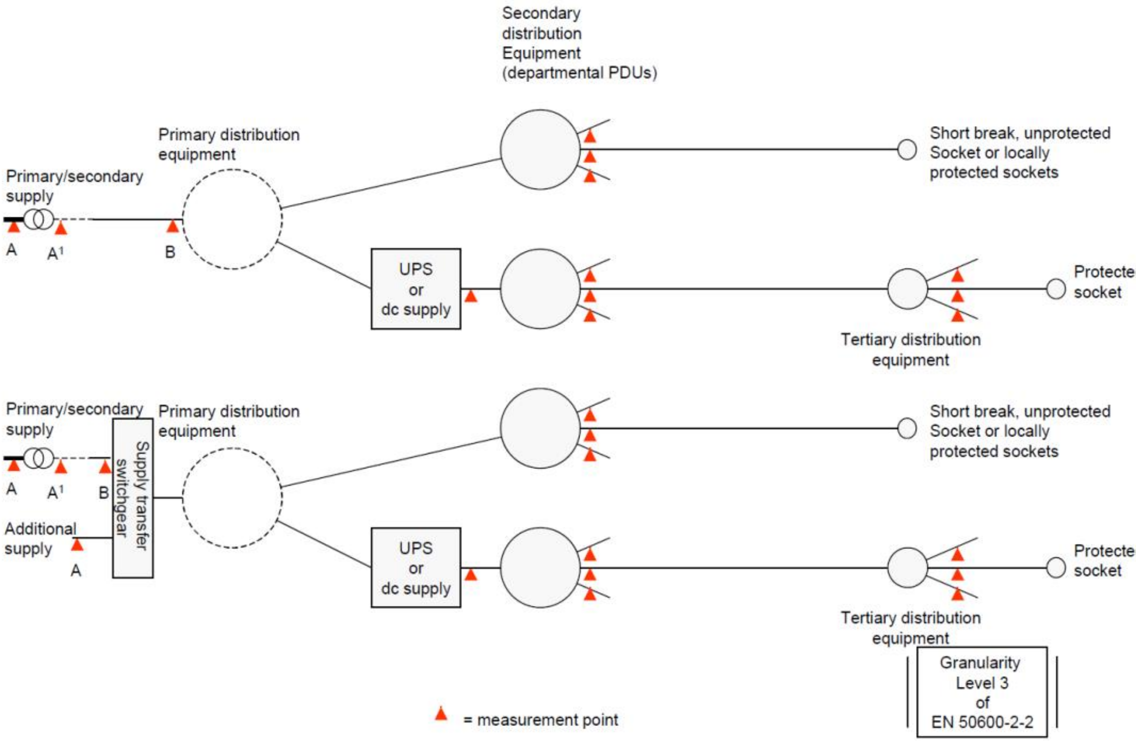

Secondary distribution Equipment (departmental PDUs)
Primary distribution equipment
Primary/secondary supply
A
A¹
B
UPS or dc supply
Short break, unprotected Socket or locally protected sockets
Protected socket
Tertiary distribution equipment
Primary/secondary supply
Primary distribution equipment
A
A¹
B
Supply transfer switchgear
Additional supply
A
UPS or dc supply
Short break, unprotected Socket or locally protected sockets
Protected socket
Tertiary distribution equipment
Granularity Level 3 of EN 50600-2-2
= measurement point

Figure 12 : Measuring points established in ETSI ES 205 200-2-1 (ETSI, 2014)

In order to cover both the ITAEI and DCFE these measurement points should collect power (Watt) and at low level, also energy data (kWh). For the DcIC power data at the equipment level is an absolute

necessity. In order to cope with the data volume and the need for timed measurements, all the measurement points should be reachable via a network connection.
While technically feasible to design a data centre that complies to the level 3 granularity as shown in figure 12, the average data centre in operation today is very far removed from this ideal measurement scenario. The reason for this lack of granularity is that the basic need of a data centre for measurements in the electrical distribution network is safeguarding the distribution network itself. Most safeguarding can be done by circuit breakers that are installed to prevent overloading on any given branch of the network, the highest level of breakers is commonly in the IT rack PDU's.

PDU's are a substantial investment, and sophisticated PDU's are even more expensive, the distribution of measurement points therefor follows the economic need for measurement and control.
The first level of sophistication in PDU's is that the PDU has a display that shows total input current in Ampere (A) this is needed to prevent connecting to much load to the PDU and when the distribution network has failover capabilities in order to safeguard against overload when load from a failed distribution branch is shifted to the surviving distribution branch. In such redundant distribution networks PDU's must not be loaded over 50% of maximum capacity.
Where needed, additional functionality in the form of a network connection is added so that input current can be monitored from a distance, but a substantial amount of data centres run without networked PDU's.
An additional complicating factor has to do with reporting accuracy. Most PDU's provide amperage data with single decimal accuracy. At 230V, 0.1 A equates to 23 Watt (assuming power factor equals 1) while this is more then sufficient for most cases, it is often insufficient to reveal differences caused by workload variations (see for example **figure 6**).
Finally, virtually none of the PDU's in use today provide monitoring at socket level making it impossible to collect the required power data from individual devices.

The problem, lack of metering on the electric socket level, can be solved by involvement of the IT equipment owner. As per the definitions given in the first section of this document, this IT equipment owner can, and in the case of Colocation data centres is, different from the data centre owner/operator (see paragraph 2.5)
Almost all IT equipment is equipped with sensors that measure the power draw of the equipment. This data can be accessed by a system administrator. The sensors invariably deliver instantaneous power readings in Watt. Unfortunately, as with many other parameters, having the capability of measurement does not guarantee that such sensors are monitored, that data is collected and stored, and most certainly does not imply availability of such data for any other party. These data types are most often used to trigger alarms. For example, zero power would indicate a fault, either a broken power supply or cable disconnection. Nonetheless, the sensors are there and software exists that can read and store this data.

The secondary need of a data centre for measurements in the electrical distribution network is billing. When a data centre has clients that are billed for their energy use, PDU's that contain a kWh metering are used. Together with the power metering and network connection, these PDU's can report the total energy and power used by the IT equipment in the rack at a user chosen interval. Recording of the total data centre energy use is usually available because the DC has to pay the utility provider.

From an electrical data standpoint, both the ITAEI and DCFE are very flexible. The required data for the ITAEI is "average power of IT" and DCFE requires only "total data centre energy use". As such, the level 3 granularity is not absolutely required for these metrics. In essence, measurement points are needed only at:

- Energy metering at primary distribution level (total DC energy)

- Energy metering at secondary distribution level (total IT energy)

Most data centres collect this data because this data is also needed for the commonly used PUE metric. Where such data is not collected, mandating installation of metering is reasonable because of the limited number of meters needed to collect the requested data.
The average IT power over an interval needed for calculating the ITAEI can be obtained easily by taking the energy draw over the interval and divide it by interval length.
As with PUE measurements, determining IT power draw at the secondary distribution level introduces a systematic error, all distribution losses are counted towards IT energy use and this creates the illusion of efficiency.

Conclusion on the quality and availability of electrical data:
The definition of the ITAEI and DCFE metrics allows for a relatively easy and good quality electric data collection. The quality is influenced by the measurement granularity but even when conducted at the secondary distribution level, the systematic errors (due to distribution losses) are in general smaller then 3%.
For the DcIC and underlying SIC, there is need to acquire data at the IT equipment level. However, as will be discussed below, this data can be accumulated by the equipment owner and only the idle energy needs to be published. The availability of such data for the authorities remains a separate question but for calculation of the proposed metrics, it is safe to conclude that this data is available to the data centre and of good quality.

IT usage data

The collection of data on IT usage is far more complex then the collection of electrical data.
The afore mentioned document (EDNA, 2022) prescribes a measurement methodology:
ITAEI is a metric based on an average of the results of testing used capacity and power over a sampling window, i.e., a representative period of time, hence it does not measure total work delivered with reference to the total energy consumption in that period of time. Testing is proposed to follow the recommendations from the Green Grid. The testing time would be between one to ten minutes aggregated into an average of two hours. Sampling windows should include peak bursts within each four-hour window per day during a 30-day interval.

Several factors complicate matters:

- Number of data sources.
  Where a fairly limited number of data sources suffice in the electrical network, all IT sources must be accessed in order to obtain IT usage data. Taking a reasonable value for the power draw of an average piece of IT equipment, 250 W, a 10 MW data centre could hold up to 40.000 units in a mix of servers storage and network equipment.
- Data volume
  The number of sources must be multiplied by the number of samples from each piece of equipment. At least 1 sample per 10 minutes in 4 hour windows for 30 days equals 720 samples per measurement point.
- Time, according to the prescribed measurement methodology, all measurements must be conducted in the same 4-hour window

The challenges mentioned above are solvable by deploying automation. A suitable software package could poll agents on each piece of equipment and collect data in a database for further analysis.

It must be noted that much IT equipment is under the supervision of some form of management software, however, like mentioned in the previous paragraph, most of these packages do not record data, but show only alarm status. Not actual usage, but an alarm when a threshold is crossed. Collecting data would require a different setup. This is not impossible, but simply not current practice.

Issues arise also in the determination of server capacity per Watt and determination of the used server capacity. The metrics document proposes:

TOPS is a unit of "Composite Theoretical Performance" (CTP), which is a measure of the performance of computers used in international agreements and Export Control in Japan. In the case of a single processor machine, CTP measures the execution speed of arithmetic and logical operations and converts the execution speed relative to the word length to 64-bit word length. In the case of multiprocessor machines, the above calculations are performed for each processor, and the weight coefficients are multiplied to the results and the sum is taken. GTOPS (= 10^6 TOPS) is often used as a unit.

A complication arises from the increased use of GPU, ASICs and other heterogeneous computational elements combined with the variations in core frequencies that are common in modern servers. Determination of the arithmetic performance becomes highly dependant on the application type and whether or not accelerators are used in the calculations.

At this point, less ambiguity exists in the definition of network and storage usage, we do suggest the binary notation: 1 kB = 1024 Bytes, 1MB = 1024 kB etc.

The notation for total usage might be somewhat confusing for static data types. For storage, total usage is expressed in storage hours. So, if for example 1 MB of data is kept for 24 hours total usage would be 24 MB.h

Technological issues will remain to be solved, for example data stored on non volatile media (NVRAM but also tape media) does not use energy when it is not accessed. Such data can have an unwanted influence on the determination of the metrics.

Two additional factors also pose a serious threat for data availability:

The first factor is organisational, valid especially for co-location providers. IT System management is distributed over multiple organisations, in the case of co-location, over possibly thousands of customers. The willingness to cooperate (share data) and technical prowess (ability to collect data) of these customers cannot be ascertained. Recent experiences of the author with the collection of data on power management by co-location customers also highlighted that there are legal issues involved when trying to enforce an obligation on a co-location data centre to provide data that is owned by a co-location customer.

Coordinating the timing of data acquisition by a large (or any) number of customers should be deemed impossible. Even if such an attempt were made, data quality can not be assured.

The second serious issue is technical and concerns the developments in software defined infrastructures and the developments in virtualization as discussed in the data centre trends section.

The simple designation of the labels “server”, “storage” or “networking” equipment is no longer maintainable. Servers can and do function (partially) as storage or network equipment. Due to the software defined storage structure, there is no real distinction between an internal disk in a server or a multi TB storage subsystem somewhere in the network. The question thus becomes, where to do we assign the power draw of an internal disk, and/or its storage capacity?

Networking poses even larger issues, since switching is incorporated in the computational clusters, through hardware when deploying blade systems or software when deploying virtual switching, the question arises “where do we measure network activity?”. Looking at physical/virtual port level, which would in theory provide the volume of all network traffic is complicated because with traffic becoming more and more concentrated to internal (within a server) virtual switches, the number of network ports to be monitored becomes astronomical.

For the determination of the DcIC, it is necessary to determine the energy used for “idle”. The document “the idle coefficients” describes how this is accomplished, in a nutshell however, each server inside the data centre should collect both electrical power draw and CPU utilization at regular intervals.
From this data the total energy use and the idle energy use can be calculated:

$$\boldsymbol{Server\ Total\ Energy} = \sum_{\mathbf{1}}^{\boldsymbol{N}} \boldsymbol{P}(\boldsymbol{n}).\,\boldsymbol{t}(\boldsymbol{n})$$

Where

- n is the time interval number;
- P(n) the recorded power for interval (n);
- t(n) the duration of the interval (n).

Determining the idle energy involves both the CPU idle **time** as well as the **power draw** of the server when idle. Most monitoring software solutions record/provide CPU load, but not idle. Therefore, the CPU idle energy needs to be calculated with a difference to a 100% load (100%- CPU load%); which results in the following formula:

$$\boldsymbol{Server\ Idle\ Energy} = \sum_{\mathbf{1}}^{\boldsymbol{N}} [\mathbf{100}\% - \boldsymbol{CPU}\%(\boldsymbol{n})].\,\boldsymbol{P}_{\boldsymbol{idle}}\,.\,\boldsymbol{t}(\boldsymbol{n})$$

Where

- n is the interval number;
- CPU%(n) the recorded CPU load (%) for interval n;
- P_{idle} is the idle power of the server;
- t(n) is the duration of the interval (n).

The duration of the interval is flexible but should reflect the time between changes in workload. The intervals do not have to be equal for different servers nor does it have to be constant for a single server. The calculation itself is trivial, for purposes of data reduction, the calculation can be done on the server itself, transferring only the total and idle energy to a central collection point.
The DcIC is calculated by a simple summation of all the reported energy data.
The final calculation ensures total anonymity for the IT equipment owner. Not only individual usage data, but even the total energy usage is removed since the resulting metric is only a ratio.

As with the ATAEI and DCFE, the organisational aspects also affect the DcIC, especially for co-location providers, IT System management is distributed over possibly thousands of customers. The willingness to cooperate (share data) and technical prowess (ability to collect data) of these customers cannot be ascertained.
Technical issues are less of an issue. The DcIC relies on the most commonly monitored sensors: CPU load and electrical power only. Since the idle coefficients are only concerned with idle cycles, the aforementioned issue of functional blurring is also no issue. The idle coefficients are determined irrespective of the function that a server performs.

Conclusions on IT usage data:
In order to calculate the proposed ITAEI and DCFE, a clearer definition of what is labelled “storage, networking and server equipment” is needed and a precise description of what and where to measure must be added.

At this point in time it is safe to assume that, with the occasional exception, IT usage data needed for the calculation of these metrics is unavailable. The type of data is at best monitored for threshold exceedance, but not collected/stored/analysed.

Even the future availability of such data within the confines of a data centre, is questionable, partially because of technical, but also organisational and confidentiality barriers. Additionally, the cost and effort involved in the collection and handling of the expected volume of data is also a barrier for adoption.

With consideration for the loss of the detailed information contained within the ITAEI and DCFE, the DcIC can be considered to be a useful alternative metric. The availability of data and the limited volume of this data can foster the adaptation of this metric. When combined with the already widely accepted PUE, de data can provide a single number that indicates the potential for further energy savings within the data centre. This combined metric can not be coined a “functional efficiency metric” but even without the direct link to “useful work” can be used to track improvements in both the data centre infrastructure energy efficiency as well as the operational efficiency of IT equipment.

5 Conclusions and recommendations

This document has described a simple and straightforward definition and categorization of data centres usable for policy creation.

Definition: A Data centre is a structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of, and including, 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.

The number of categories proposed are purposely limited and are used only to assign responsibility for the management of certain aspects of the data centre on which policies might be targeted.
These aspects are;

- Operation of the building infrastructure
- Operation of the IT equipment
- Operation of the software.

Only three categories are therefor proposed:

- *Enterprise data centre*: the building, hardware and software all have the same owner/operator (the data centre owner).
- *Co-hosting data centre*: the building and the hardware are provided by the data centre operator/owner (the co-hosting provider), (most of) the software is operated by (multiple) co-hosting customers (lessees).
- *Co-location data centre*: the building is provided by the data centre operator/owner (the co-location provider), (multiple) co-location customers provide their own hardware (servers, storage and network equipment) and software.

Even with such a straightforward definition and categorization we find that the availability and quality of data sources for trivial data such as location and ownership of data centres is low.

Where authorities are aware of the existence of data centres within their region, data on energy use are often unavailable since these are considered to be company (competition) sensitive (Brocklehurst, 2021).
This data even when not easily accessible, is technically available. Every DC location is connected to an energy grid and data networks. Energy bills are paid and equipment installed and serviced. For this, ownership, location, energy use, available capacity etc. is all information that is recorded and kept. Where the ownership of the site infrastructure systems differs from its user(s), customers are billed for the use of the data centre and as such their information is also recorded.

For data needed for the determination of data centre efficiency, for which the metrics IT Equipment Average Efficiency Index (ITAEI) and Data Centre Functional Efficiency (DCFE) as well as for the non-functional metric, the DcIC, are proposed, availability is much worse. No such data is publicly available. This document therefor focussed on the technical availability or obtainability of this data.

The electrical data needed for the calculation of these metrics is technically available since most data centres have excellent management and control over this element in the “operation of the building infrastructure”. Although few data centres have the granularity of measurement points suggested for the determination of the metrics, most have a usable level of measurement granularity which only introduces a manageable systematic error into the metrics.
Where there are insufficient measurement points to acquire the needed data, obligating installation of the minimally required energy management infrastructure is reasonable since the number of meters needed is comparatively limited.

IT usage data is currently unavailable, it is very rarely collected. With the caveat of proper definition, obtaining such data is technically feasible, but the effort associated with collection of much of this data is expected to be very high.
Especially true for the functional metrics ITAEI and DCFE, the number of measurement points for correct determination in a large data centre depends on the exact definition of the needed data but can range from tens to hundreds of thousands resulting in data volumes of hundreds of millions of data points.
A complete description of both what and where to measure needs to be provided to assure any type of data quality on IT usage, but such a description might require a much more complicated second dimension in the data centre categorization which would describe the type of work and consequently type of IT infrastructure present in the data centre.
Since IT infrastructures are designed specifically for their intended use, the current three data centre categories might need to be augmented when these metrics are fully adopted in order to collect the proper statistics and compare with a relevant “best available technology”.
For exmple

- The basis for server selection and therefor usage statistics can be memory capacity rather then CPU
- The basis for storage selections could be IOPS not GB
- Network criteria could be latency rather then bandwidth.

Unwanted influence of the metrics on equipment selection and IT architecture, could be remedied by adjusting the “best available technology” used as reference and/or the capacity unit used in the metric determination, for data centres with different workloads. This would however severely complicate the categorisation of the data centres.

The use of a non-functional metrics resolves many of the data issues that are associated with the functional metrics. The DcIC, a non-functional metric, needs only two data types from each server, electrical power and CPU utilization. When data volume is an issue, the first step in data processing can be done by the server itself, limiting the data flow to just 2 data points per server. With the ease of determination however comes a loss of information. The idle coefficient does not consider dedicated storage and networking equipment and is not sensitive for IT equipment renewal. Nonetheless, as the PUE has demonstrated, the optimization of an imperfect indicator can still result in marked improvement in energy efficiency of data centres.

Overall, a phased approach is recommended for any policy creation towards the data centre sector. As proposed by the energy efficiency directive (EED) of the European Union, the first step would be to acquire basic information by requiring data centres to register and update seemingly trivial information

- Contact information
- Location(s) and category (as per the proposed definition)
- Maximum power draw and actual energy usage

If required, such data can be collected under some form of non-disclosure between the authorities and the parties involved, but a complete and accurate map of the data centre landscape must be available to policy makers.

Secondly, since utilisation is such a determining factor in efficiency, the sector (representatives) should be involved in preparing the adoption of a suitable efficiency metric.

REFERENCES

ADEME. (2022, 11 1). Retrieved from ADEME home page: https://www.ademe.fr/en/frontpage/

ASHRAE TC9.9. (2022, 06 29). *Mission Critical Facilities, Data Centers, Technology Spaces and Electronic Equipment.* Retrieved from https://tpc.ashrae.org/?cmtKey=fd4a4ee6-96a3-4f61-8b85-43418dfa988d

ASML. (2022, 08). *EUV lithography systems*. Retrieved from https://www.asml.com/en/products/euv-lithography-systems

Avgerinou, M., Bertoldi, P., & Castellazzi, L. (2017). Trends in Data Centre Energy Consumption under the european code of conduct for datacenter energy efficiency. *Energies*.

AWS. (2022, 07 07). *What is Amazon EC2*. Retrieved from https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

Bauer, D., & Papp, K. (2009). Book review perspectives: The Jevons paradox and the myth of resource efficiency improvements. *Sustainability: Science, Practice & policy 5 (1).* doi:doi:10.1080/15487733.2009.11908028

Bizo, D. (2019). *The Carbon Reduction Opportunity of Moving to Amazon Web Services.* Retrieved from https://d39w7f4ix9f5s9.cloudfront.net/e3/79/42bf75c94c279c67d777f002051f/carbon-reduction-opportunity-of-moving-to-aws.pdf

Bizo, D. (2021, 09). *New ASHRAE guidelines challenge efficiency drive*. Retrieved from DCD news: https://www.datacenterdynamics.com/en/opinions/new-ashrae-guidelines-challenge-efficiency-drive/

Borderstep institute. (2020). *Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud market.* Luxembourg: Publications Office of the European Union. doi:doi:10.2759/3320

Brocklehurst, F. (2021). *International Review of Energy Efficiency in Data Centres.* the Australian Department of Industry, Science, Energy and Resources.

CENELEC. (2019). *EN 50600-1 Information technology - Data centre facilities and infrastructures - Part 1: General concepts.* Brussels: Cenelec.

CLC/TC 215. (2021, 08). *Information technology - Data centre facilities and infrastructures.* Retrieved from Part 99-1: Recommended practices for energy management : https://standards.iteh.ai/catalog/standards/clc/e73bdd06-fd72-4652-add9-8b34f3f0c324/clc-tr-50600-99-1-2021

D.H.Harryvan. (2021). *The Idle Coefficients.* 4E-EDNA. Retrieved from https://www.iea-4e.org/wp-content/uploads/2021/10/Server-Idle-Coefficients-FINAL-1.pdf

EDNA. (2022). *Metrics for Data Centre Efficiency.*

EDNA. (2022, 10 20). *task-28-data-centres-workstream-activity-1*. Retrieved from EDNA tasks: https://www.iea-4e.org/edna/tasks/task-28-data-centres-workstream-activity-1-definition-of-scope-trends-and-data-availability-and-quality/

ENERGY STAR. (2022, 06 29). *portfoliomanager glossary*. Retrieved from https://portfoliomanager.energystar.gov/pm/glossary#DataCenter

engineering toolbox. (2022, 06). *pump affinity laws*. Retrieved 6 2018, from engineering toolbox: https://www.engineeringtoolbox.com/affinity-laws-d_408.html

ETSI. (2014, 03). *ETSI ES 205 200-2-1*. Retrieved from https://www.etsi.org/deliver/etsi_es/205200_205299/2052000201/01.02.01_60/es_2052000201v010201p.pdf

EU-JRC. (2022, 6 13). *Code of Conduct Data centres.* Retrieved from Data Centres Energy Efficiency: https://e3p.jrc.ec.europa.eu/communities/data-centres-code-conduct

eversheds-sutherland. (2022, 11 2). *German government considers energy efficiency, renewable energy and disclosure requirements for data centers from 2023 onwards*. Retrieved from eversheds-sutherland: https://www.eversheds-sutherland.com/global/en/what/articles/index.page?ArticleID=en/Energy/German_government_considers_energy_efficiency_renewable_energy

General Secretariat of the Council of the european union. (2022, 06 24). *Proposal for a Directive of the European Parliament and of the Council onenergy efficiency (recast).* Retrieved from https://data.consilium.europa.eu/doc/document/ST-10490-2022-INIT/en/pdf

Heslin, K. (2015, 09 14). *Switzerland's First Tier IV Certified Facility Achieves Energy Efficiency*. Retrieved from uptime institute journal: https://journal.uptimeinstitute.com/switzerlands-first-tier-iv-certified-facility-achieves-energy-efficiency/

HPCwire. (2017, 4). Retrieved from https://www.hpcwire.com/2017/04/12/hyperscalers-emerging-hype-phase/

IEA. (2022, 10 26). *global-data-centre-energy-demand-by-data-centre-type.* Retrieved from IEA.org: https://www.iea.org/data-and-statistics/charts/global-data-centre-energy-demand-by-data-centre-type

IRDS. (2022, 07 29). *what is IRDS*. Retrieved from https://irds.ieee.org/

ISO/IEC JTC 1/SC 39. (2016, 4). *Information technology — Data centres — Key performance indicators — Part 2: Power usage effectiveness (PUE).* Retrieved 2016, from 30134-2: https://www.iso.org/

ISO/IEC JTC 1/SC 39. (2021, 08). *Information technology — Data centres key performance indicators — Part 6: Energy Reuse Factor (ERF).* Retrieved from 30134-6: https://www.iso.org/standard/71717.html

Johann Schleier-Smith, e. (2021). What serverless computing is and should become: the next phase of cloud computing. *Communications of the ACM 64(5)*, 76 - 84.

Karamouzis F., A. W. (2022, 05 11). TechWave: A Gartner Podcast for IT Leaders. *Hyperscale Vendors — Too Big to Fail*. Gartner.

Koomey, J. G. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing . *IEEE Annals of the History of Computing*, 39.

Leopold, G. (2017, 04 14). *Hyperscalers Emerging From 'Hype Phase'*. Retrieved from HPC wire: https://www.hpcwire.com/2017/04/12/hyperscalers-emerging-hype-phase/

Montevecchi, F. S. (2020). *Energy-efficient Cloud Computing Technologies and Policies for an Eco-friendly Cloud Market. Final Study Report.* Vienna.

OPERAT. (2022, 11 9). *home*. Retrieved from OPERAT: https://operat.ademe.fr/#/public/home

oura, P. &. (2016). Energy savings potential of uninterruptible power supplies in European Union. *Energy Efficiency 9*.

Schödwell, B. a. (2013). Data center green performance measurement: State of the art and open research challenges. *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime*, (pp. 1003-1017).

Shalf, J. (2020). The future of computing beyond Moore's Law. *Philosophical Transactions of the Royal society A: Mathematical, Physical and engineering Sciences 378*.

Standard Performance Evaluation Corporation. (2022, 07 20). *SPECpower_ssj2008 Results*. Retrieved from SPEC's benchmarks: https://www.spec.org/power_ssj2008/results/

Synergy group. (2022, 10). *about Synergy group*. Retrieved from https://www.srgresearch.com/about

Thorsen, J. E. (2018). *"Progression of District Heating – 1st to 4th generation.* Aalborg University, Denmark.

TIA. (2017, 06). *Telecommunications Infrastructure Standard for Data Centers TIA-942-B.* TIAonline.org. Retrieved from https://tiaonline.org/standard/tia-942/

Uptime Institute. (2022, 06). *The Uptime institute tier standard: topology.* Retrieved from Uptime institute: https://uptimeinstitute.com/resources/asset/tier-standard-topology

Whitehead, B. &. (2015). The life cycle assessment of a UK data centre. *The International Journal of Life Cycle Assessment. 20. 10.1007/s11367-014-0838-7.*