LUNDS UNIVERSITET
Lunds Tekniska Högskola

Demand-response potential of large-scale data centers in Europe 2030

Carolina Koronen

Master thesis

May 2018

Dokumentutgivare, Dokumentet kan erhållas från

LUNDS TEKNISKA HÖGSKOLA
vid Lunds universitet
Institutionen för teknik och samhälle
Miljö- och energisystem
Box 118
221 00 Lund
Telefon: 046-222 00 00
Telefax: 046-222 86 44

Dokumentnamn

Examensarbete

Utgivningsdatum

2018-05-31

Författare

Carolina Koronen

Dokumenttitel och undertitel

Potential för efterfrågestyrning av storskaliga datacenter i Europa 2030

Sammandrag

Storskaliga datacenter utgör en ny energiintensiv industri. Detta examensarbete undersöker datacenters kvalitativa potential att delta i efterfrågestyrning och gör en uppskattning av vilken i vilken storleksordning denna kraftreserv skulle kunna vara i Europa år 2030. En sammanställning av forskning om strategier för grön laststyrning visar att driften av datacenter kan utformas så att elförbrukningen i största möjliga mån sker när förnyelsebar el finns att tillgå. Sådan laststyrning skulle kunna utnyttjas för att erbjuda tjänster inom efterfrågestyrning till elnätet. Med utgångspunkt i existerande scenarion för datacenters totala energibehov uppskattas potentialen för efterfrågestryning av datacenter uppgå till omkring 20-40 GW i Europa 2030. Detta innebär att datacenter tillhör några av de mest lovande sektorerna när det kommer till att erbjuda efterfrågestyrning i Europas framtida elkraftsystem. Med anledning av denna stora potential och deras snabba ökande i antal är det nödvändigt att betrakta storskaliga datacenter som de viktiga komponenter de är i morgondagens energisystem.

Nyckelord

Datacenter, efterfrågestyrning, 2030

Sidomfång	Språk	ISRN
51	Engelska	LUTFD2/TFEM—18/5129--SE + (1-51)

Organisation, The document can be obtained through

LUND UNIVERSITY
Department of Technology and Society
Environmental and Energy Systems Studies
Box 118
SE - 221 00 Lund, Sweden
Telephone: int+46 46-222 00 00
Telefax: int+46 46-222 86 44

Type of document

Master thesis

Date of issue

2018-05-31

Authors

Carolina Koronen

Title and subtitle

Demand-response potential of large-scale data centers in Europe 2030

Abstract

Large-scale data centers are a new energy intensive industry. In this master thesis, the qualitative potential of engaging data centers in future demand-response schemes is assessed, and the magnitude of the power reserve available for this purpose is estimated for Europe 2030. A research survey of green workload management strategies shows that data centers can be operated in a way that aligns their power demand with renewable energy availability. Such workload management could be leveraged to provide demand-response services to the grid. Based on adopted scenarios of data center energy demand, the total demand-response potential of data centers in Europe 2030 is estimated to 20-40 GW. This makes data centers one of the most promising demand-response providers in Europe's future power system. Considering this potential and their rapid increase in numbers, large-scale data centers must be recognised and considered as important features in the future energy system.

Keywords

Data centers, demand-response, 2030

Number of pages	Language	ISRN
51	English	LUTFD2/TFEM—18/5129--SE + (1-51)

Demand-response potential of large-scale data centers in Europe 2030

Carolina Koronen

Acknowledgements

I would like to express a special thank you to

Max Åhman and Lars J Nilsson for introducing me to the subject of data centers and giving me the opportunity and freedom to thoroughly explore it. Furthermore, to Max for sharing your time, knowledge and encouragement throughout the process of this master thesis in your role as supervisor.

Matilda Axelson and Tomas Wyns for your support and input during my work from Brussels. Also the many other friends and colleagues at Institute for European Studies at Vrije Universiteit Brussel who have given me an inspiring and friendly workplace that I have enjoyed going to every morning.

Jesse Pappers and my family and friends, who may not feel like they contributed much to this thesis but without whom it would never have been finalised.

EXECUTIVE SUMMARY

This thesis assesses the growing data center industry and its potential contribution in demand-response in 2030's Europe.

Data centers are facilities containing servers that host various types of IT-services. Small data centers, called server closets, have a power draw of a few kilowatts and are commonly found inside office buildings. Their highly specialised large-scale counterparts have a power draw of up to hundreds of megawatts and include a highly efficient cooling system and an internal power system with back-up generators. Large-scale data centers are more energy efficient than smaller ones, and their servers are more often shared between many users through cloud computing.

The demand for digital services and for computing capacity is growing, and as more businesses and consumers turn to cloud services instead of own computer capacity to satisfy this demand, large-scale cloud data centers are expected to increase in numbers and replace many of the server closets. Their large electricity demand, along with an inherently variable productivity level caused by a close relationship to human rhythm of day, make large-scale data centers interesting candidates for demand-response.

Data centers have a qualitative potential to engage in demand-response. Previous research surveyed in this thesis show that data center operators can achieve a larger share of the green energy in their electricity mix by prioritising to run workloads (computing jobs) when and where the availability of renewable energy is high. Data centers handle different types of workloads, of which some have a certain delay tolerance. This delay tolerance can be leveraged to schedule the workload to run at the most suitable time within its pre-defined deadline. Clusters of data centers with duplicated data and applications also have the ability to choose in what data center the workload is handled. This type of workload management could be applied with the objective of providing demand-response services to the grid.

The demand-response potential is of considerable magnitude. In this thesis, the potential has been estimated using adopted energy scenarios for data centers and qualified assumptions of data center operational constraints. Based on the assumption that the energy demand of the European data center fleet in 2030 amounts to 200TWh, the demand-response potential is estimated to around 20-40GW. This is comparable to two other promising demand-response providers electric vehicles and commercial buildings and could lessen the need for battery banks and transmission capacity in the 2030 European power system.

To leverage the demand-response potential, a financial incentive for data center operators is needed. This could be created by establishing a functioning demand-response market. In spite of their promising demand-response potential and their considerable future electricity demand, data centers have yet to be viewed strategically from an energy system point of view. To ensure a sustainable and resource efficient emergence of this new energy intensive industry, data centers need to be recognised and considered as important features of the future energy system.

CONTENTS

1 INTRODUCTION

This chapter introduces the topic of this master thesis with a brief background on renewable energy transition and digitalisation, before defining the objective and scope of the study, and describing the method and disposition of this report.

1.1 Background

Mitigating climate change is possibly the greatest challenge for humankind in the 21st century. Keeping the Earth's temperature increase below 2°C, as agreed in the Paris Agreement, will require transitions to low-carbon technologies in almost every sector – and most importantly in the generation of electrical energy. Meanwhile, another technological revolution is happening. Gathering and processing of digital information has become tremendously easy and cheap, resulting in new data-driven applications to be launched every day. A digital industry of data centers – specialised data storage and processing facilities – are rapidly growing in numbers and in demand for electricity.

"So, we have two parallelly ongoing technological and social developments with a common emphasis on electrical energy supply – could there be synergies to explore between the them? How can these technologies co-evolve for increased resource efficiency?" This thought sparked this master thesis.

1.1.1 A new energy system

The world's energy systems are facing a shift of paradigm. Over the past hundred years, humanity has made fast technological, social and economic progress, but a prerequisite for almost all of it has been the access to inexpensive energy from the burning of fossil fuels. In the process of creating welfare, humans have caused a drastic rise in temperature on Earth, which now threatens to push the planet's physical and biological systems over their tipping points. The 21st century will be all about slowing down and bating the consequences of climate change. To do this, steering away from fossil fuels is a key issue.

Though numerous sectors of the economy are subject to mitigation efforts, the generation of electrical energy, often simply called the *power sector*, plays a very central role. Decarbonising electricity generation, i.e. replacing carbon intensive generation, such as coal-fired plants, with low-carbon technologies like solar and wind power, is a cost-effective way of cutting carbon emissions (IPCC, 2014). Electricity will also come to represent a larger share of our final electricity consumption in the future, as it takes over the role as energy carrier in other sectors, most importantly transport. (OECD/IEA, 2016)

The European Union (EU), comprising almost thirty European states, has collective climate mitigation goals for three horizons. By 2020 the EU has bound itself to reducing the greenhouse gas emissions by 20% compared to 1990-levels; increasing the share of energy consumption originating from renewable sources to 20%; and improving energy efficiency by 20% compared to projected levels. By 2030, these figures are increased to 40%, 27% and 27% respectively. For the 2050 horizon the EU is committed to an 80-95% reduction in greenhouse gas emission, compared to 1990-levels (European Union, 2018).

Policy goals aside, the cost of producing renewable electricity has been steadily decreasing over the past decade. By 2020, all currently commercial renewable power technologies will be within the same cost range or cheaper than fossil-based technologies, according to the International Renewable Energy Agency (IRENA, 2018). Because of these price drops, a fast growth of renewable installations and a replacement of fossil-based technologies are to be expected and the renewable power increase driven by this improved competitiveness is often underestimated by policy makers (Metayer *et al.*, 2015).

As a result of these driving forces, it is likely that roughly 50% of generated electricity in Europe originates from renewable sources in 2030. Wind power will be the largest contributor, followed by hydro power, solar power and biomass. (Banja and Jégard, 2017)

Balancing a highly renewable power system

A challenge with this much-needed transition to renewable energy sources is that an electrical power system must always remain in balance between supply and demand of electricity. Traditionally, these systems have been operated in a supply-following manner, where balance has been achieved by altering the production level in the power plants, to satisfy the demand. Such a scheme requires *dispatchable* production, i.e. power plants whose power output can be controlled. Two of the dominating renewable energy technologies, namely solar and wind power, are *non-dispatchable,* meaning that their power output cannot be controlled. The increased share of renewable energy sources therefore calls for structural changes in the power system infrastructure, operations and markets (OECD/IEA, 2016).

There are essentially three types of measures that, in combination, can create stability in a high-renewable power system: Storage, transmission and demand-response. Storage can smooth out the variations in time, by storing energy in periods of high supply and low demand and using it when circumstances are the opposite. Transmission overcomes spatial mismatch, by overbridging the often-long distances between large renewable production sites and areas with high demand. While storage and transmission are measures that still works towards creating power availability that meets consumer's power demand, demand-response is about re-shaping that demand. In demand-response schemes, consumers adjust their electricity consumption according to variations in the local grid. This can be incentivised through electricity price or regulated in an agreement with the grid operator. Both individuals and large industries are potential participants in demand-response. This brings us to the next technological development.

1.1.2 A new energy intensive industry

Society is increasingly data-driven. The emerging Internet of Things (IoT) and its ubiquitous gathering of information, could soon enough make data-analytics part of almost every business (Cisco, 2014). Also on the consumer side, data-intensive activities like video streaming and social networking are growing in popularity. In fact, it seems that nearly all of today's buzz words, from smart cities and artificial intelligence (AI) to big data and cryptocurrency, are about data and information and communication technology (ICT).

While user devices like smartphones and laptops are the visible part of ICT-industry, an increasing share of the actual digital services takes place in data centers. In the last two decades or so, the world's data center fleet has grown from a few enterprise computing centrals to an important feature in modern infrastructure (Corcoran and Andrae, 2013). In 2015, about 3.3 million data centers existed worldwide (European Commission, 2015), handling internet traffic and data of astronomical numbers – 4.7 ZB of IP traffic and 171 EB of stored data[1]. By 2020 the traffic rates are expected to triple and the amount of data stored to increase five-fold (Cisco, 2014). In short: This is industry is booming.

Both data centers, networks and the user devices' operational phase are substantial contributors to the total energy demand of ICT industry. In 2012, the operational phase of the devices (e.g. the electricity used to charge your phone) was the largest, at almost half of the power consumption. While this share is shrinking, the contribution from data centers and networks is increasing. (Corcoran and Andrae, 2013)

The electricity demand of data centers already represents several percentages of that of society as a whole (Koomey, 2011; Andrae and Edler, 2015) and many expect it to keep growing in response to the increased use of data.

Data centers and demand-response

Data centers interesting from a demand-response point of view. They are large consumers of electricity, who are growing in numbers, and their operations are already highly connected, monitored and

[1] 1 zetabyte (ZB) = 10^{21} bytes, 1 exabyte (EB) = 10^{18} bytes

responsive to control signals. Furthermore, many IT-services have varying activity levels throughout the day and in some cases a certain delay-tolerance. These characteristics could possibly be leveraged in demand-response programmes and facilitate operations of a future high-renewable electricity grid.

1.2 Objective

The purpose of this study is to assess the growing fleet of large-scale data centers and analyse if and how they could contribute to balancing a highly renewable European power system in 2030 and beyond.

1.2.1 Research questions

More specifically, this work will focus on the following questions:

- How can large-scale data centers be leveraged in demand-response?
- How much of data centers' power demand could be available for demand-response in Europe 2030?

To be able to answer them in a comprehensive way, this work will also describe data center technology with an emphasis on energy aspects.

1.3 Scope

- The focus of this study is on the technological nature of data centers and their potential power flexibility. Other aspects, such as legal restrictions, will not be analysed.
- This study focuses on data center load flexibility without involving energy storage or back-up generators.
- This work focuses primarily on the development towards the year 2030.

The year 2030 was chosen as a horizon because it is subject to concrete targets for the energy system and it is within a reasonable range for projections of the data center industry and the power system scenarios. However, it is far enough into to the future to allow time for strategic decisions and actions to be made that could change the course of development.

1.4 Method

First, a brief technology assessment of data centers and their energy properties was carried out, based on literature search in academia and industry. Characteristics of the present European data centers fleet were mapped and trends were identified. Next, to assess and demonstrate the technological opportunities for data centers engaging in demand-response, a research survey was performed, covering a selection of academic studies on green workload management strategies for data centers with different characteristics in different energy system contexts. Then, the order of magnitude of data center flexible power in 2030 was estimated, through a calculation based on energy scenarios of the future data center fleet and qualified assumptions on utilisation patterns. Finally, the findings were discussed and conclusions were drawn.

1.5 Disposition

This first chapter presented the background to this master thesis and outlined its objective, scope and method. Chapter 2 gives a comprehensive introduction to the structure, subsystems and components of data centers, as well as cloud computing terminology. The energy performance and efficiency indicators of data centers are described in chapter 3. Chapter 4 is a mapping of the European data center fleet, with respect to scale, end-users and geographical hubs. Chapter 5 deals with the flexibility opportunities of data centers by introducing the concept of demand-response and then surveys research on green workload management. Chapter 6 describes the calculation of the estimated available flexible power from data centers in 2030. The findings of the previous chapters are discussed in chapter 7.

2 DATA CENTER INDUSTRY AND TECHNOLOGY

This chapter is largely a high-level technology assessment, covering the basic infrastructure, hardware and software subsystems of data centers. It also includes a description of common data center business terminology.

2.1 What are data centers?

A data center is essentially a room with a set of network-connected computers that host IT-services, along with any supporting infrastructure. A suggested definition of data centers, formulated by the Australian Department of Industry (Commonwealth of Australia, 2014), reads as follows:

> A data centre is a structure, or group of structures, located on a single site dedicated to the centralized accommodation, interconnection, and operation of information technology and network telecommunications equipment that provides data storage, processing, and transport services. A data centre encompasses 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.

Figure 1 below is a photo from inside a data center and shows a typical organisation of servers in racks, along isles.

Figure 1: Photo from inside a data center. Source: [http://www.bloter.net/archives/150382].

2.1.1 Data center system overview

Figure 2 depicts a simplified overview of the different systems in a data center, arranged in a pyramid where the systems at the bottom support or host those above. The power and cooling infrastructure are ancillary to the IT-equipment, or hardware. The IT-equipment is installed with an operating system and virtualisation software, on which applications run. The applications provide IT-services by handling workloads. In addition, a resource management system makes decisions about workload priority and resource allocation. Each of these systems are discussed individually in the following sections.

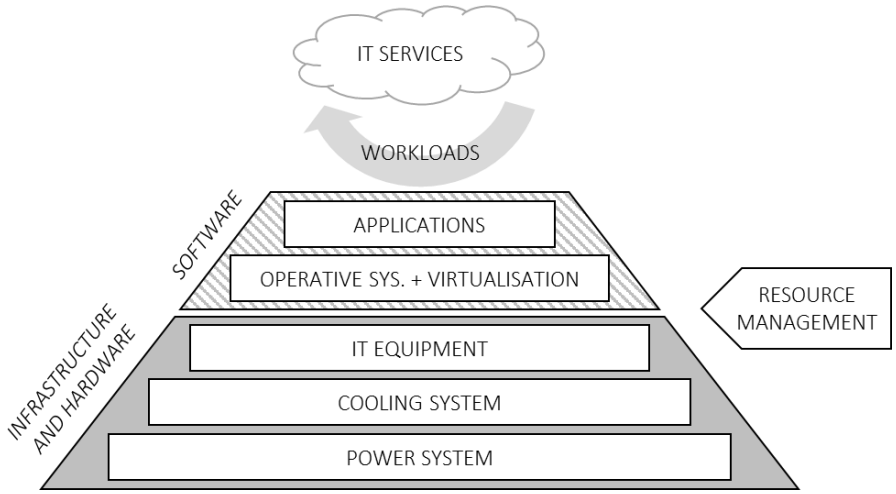

Figure 2: The data center system hierarchy.

2.2 Infrastructure and hardware

The IT-equipment in a data center includes servers, storage and network. To ensure good operations, data centers also contain supporting infrastructure, most importantly a power system and a cooling system. The industry exhibits heterogeneity regarding the design of IT-equipment as well as supporting systems but follows similar principals. Additional systems commonly found in data centers are monitoring systems to control operations and surveillance systems to ensure security, but these are omitted in this work.

2.2.1 IT-equipment

The most defining building blocks of a data center are the servers. The computing functionality of the server is given by a processor, often called CPU (Central Processing Unit), together with a chip-set and memory components. The servers also have local power supply/distribution units (PSU/PDU), that transforms the power to a suitable voltage before feeding it to the electronics, as well as network connection points, local storage and internal cooling. Figure 3 depicts a server's internal structure. (European Commission, 2015)

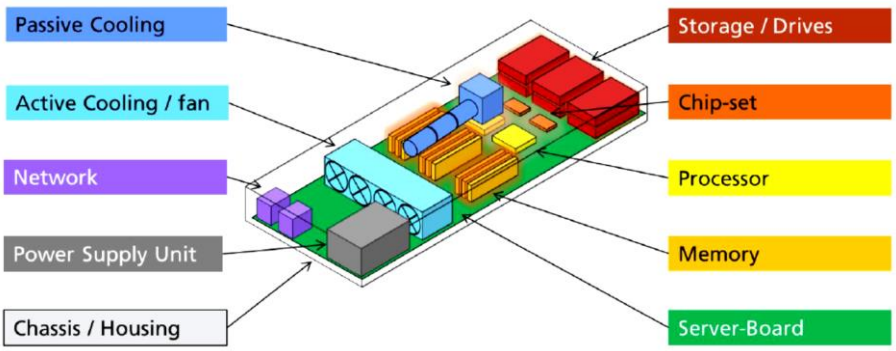

Figure 3: Schematic picture of a server's internal design. Source: (European Commission, 2015).

To complement the servers' internal memory and storage, the IT-equipment in a data center usually includes separate storage devices with larger storage capacity. This hardware consists of storage media along with power supply units, cooling and network connections, similarly to servers. (Barroso *et al.*, 2013c; European Commission, 2015)

Cables, switches and routers constitute a multi-layered network. The network connects servers and storage with each other throughout the data center and ultimately with the outside world and the clients using the data center. (Barroso *et al.*, 2013c)

The IT-equipment is chosen according to the needs of the applications they are intended to host. The components can vary in performance level (computing speed or storage capacity), scalability and physical dimensions, and service level aspects like availability and resilience. IT-equipment performance is furthermore highly dependent on the physical environment – temperature and humidity. (European Commission, 2015)

2.2.2 Power system

A reliable and high-quality power supply is important to the data center. The IT-equipment is made up of sensitive electronic components that can be damaged by voltage spikes. The IT-systems hosted in data centers also often have high requirements of uptime which can be compromised by a power outage.

For protection against fluctuations in the incoming power, data centers are typically equipped with a double power conversion mechanism which guarantees that the right voltage and frequency are fed to the facility. Power distribution units that adapt the voltage and frequency according to equipment needs are also often present at several points in the data center's electrical cabling. (Barroso *et al.*, 2013a)

For protection against outages, data centers have batteries and diesel generators that together form an *Uninterruptable Power Supply* (UPS) system. The battery power cuts in immediately upon grid failure but can only support the data center short-term. While the data center runs on battery power the diesel generators are warmed up and eventually take over as back-up power source. (Barroso *et al.*, 2013a; Commonwealth of Australia, 2014)

The back-up power system must at a minimum keep the data center running long enough to allow a safe shut-down of the servers. In many cases though, downtime is practically unacceptable. The back-up power system must then be dimensioned keep the operations running until the grid power is likely to be back. As an example, at Facebook's facilities in Luleå, Sweden the back-up power system is designed to support the data center for up to 32 hours (Länsstyrelsen Norrbotten, 2011).

2.2.3 Cooling system

Cooling is almost as crucial to operations as the power supply. As electricity flows through its circuits, the IT-equipment in the data center generates heat. If this thermal energy is not removed, servers will quickly be overheated which will cause failure. Cooling systems are therefore found in every data center (with exception for the very smallest ones) and are normally supported by the back-up power system, in case of power outage.

The specific design of the cooling system varies from one data center to another depending on the climate at the site and on the facility restrictions. The medium used to remove the heat from the servers is usually cold air, which is circulated through the racks. The resulting hot air is either re-cooled in a CRAC-unit (Computer Room Air Conditioning) and circulated back to the racks in a closed-loop system or rejected and replaced by cool outdoor air in an open-loop system. Using outdoor air, often called "free cooling", is energy efficient compared to other cooling systems but requires a relatively cold climate on the data center location. In specialised cases (primarily in high performance computing) where the media must carry larger heat load, a liquid cooling medium is used instead of air. (Barroso *et al.*, 2013a; European Commission, 2015)

For protection in case of failure, it is common for data centers to have back-up cooling equipment or to be running parallel systems where one line is underutilised during normal operations. (Commonwealth of Australia, 2014)

2.2.4 Energy system overview

Figure 4 shows a simplified schematic view of the main components in the local energy system in a data center and the energy flows between them. The blue arrows are electrical energy and the red arrows are heat. The arrows are not to scale with the size of the energy flows.

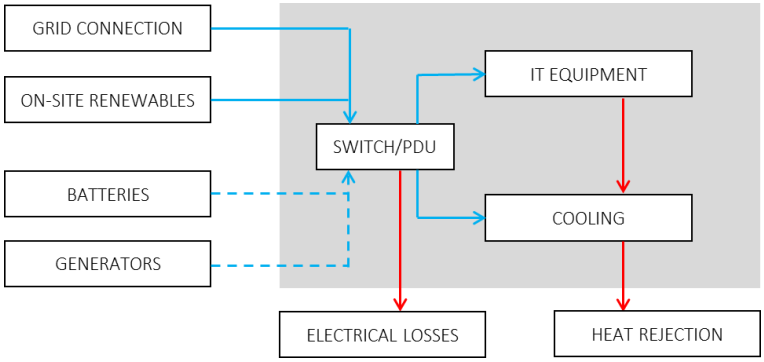

Figure 4: The data center energy system and energy flows.

2.2.5 Scalability

The scale of data centers can differ with several orders of magnitude and are given in various measures: Number of connected servers, the floor space they occupy (m^2 or ft^2), the total data center power consumption (kW) or the power consumption of the IT-equipment only (kW). The latter is called *IT-load*. Among these, IT-load is perhaps the most closely related to the data center's useful work, carried out by the servers. 1 kW of IT-load represents roughly 2 kW of total data center load in an average data center (with PUE = 2.0, see section 3.1). (European Commission, 2015)

Data centers with only a few servers and a connected IT-load of less than about 10 kW are referred to as *server closet* or *server room*; facilities with an IT-load in the range of about 10 kW to 750 kW are considered *small-* or *medium-scale* data centers; data centers with IT-loads ranging from 750 kW to 2.5 MW are considered *large-scale*; and finally, data centers with an IT-load of 2.5 MW or more are called *megascale* (Commonwealth of Australia, 2014). These ranges are summarised in Table 1. However, they are not strict definitions and different literature use slightly different terminology and ranges.

Table 1: Suggested data center scale categories by connected IT-load.

Category	**Range IT-load**
Server closets and -rooms	< 10 kW
Small-scale data centers	10 – 150 kW
Medium-scale data centers	150 kW – 750 kW
Large-scale data centers	750 kW – 2.5 MW
Megascale data centers	> 2.5 MW

Another term – *hyperscale* – is sometimes used about some large- or megascale data centers. This term is based on characteristics of the company owning the data centers, rather than the scale of the facilities. It is often reserved for data centers of a handful of the world's biggest internet companies. (Cisco, 2014; Sverdlik, 2017)

Server closets and rooms are still commonly used and often found inside the office buildings of small businesses, academia and public sector. These often lack the adequate cooling systems, power metering and sophisticated software found in their larger counterparts.

The growth in data center sector is primarily happening in the large- and megascale end of the spectrum, while server closets and rooms are decreasing (European Commission, 2015). The fact that server closets and rooms are being replaced by large- and megascale data center has positive effect on the overall

energy efficiency of data processing and contributes to the development towards a more well-defined energy intensive industry. We will return to this aspect of data center development later in this work.

2.3 Software

Data center servers run heterogenous software in a complex architecture. This section will only briefly describe some of the very basic types of software systems. The concept of virtualisation is central to the upcoming chapters and is therefore described with greater detail.

2.3.1 Applications, operating system and resource management

An *application* is the piece of software that ultimately provides a certain service, for example a streaming service or an email system. In the end, hosting applications is the whole point of building data centers. The applications run on an *operating system* that manages resources and analogous functionality. Various automated systems in a data center also keep track of operations and allocate resources. By monitoring clients' needs and available capacity or exhaust air temperature, different systems grant access to the applications or make decisions about e.g. cooling upscaling. (Barroso *et al.*, 2013d)

2.3.2 Virtualisation

Without virtualisation, each physical server hosts only one application at a time. This makes it difficult to fully utilise each server, since one application rarely needs precisely one server's capacity. One of the objectives of constructing large data centers is to efficiently share the server resources between many users and applications. Virtualisation technology is an enabler in this context.

A virtualisation software layer, as it were, "detaches" the application from the physical server, by creating *virtual machines*. The virtualisation software accesses the hardware resources in the physical servers and creates a tailor-made server structure for each application. These virtual machines can be hosted in parallel on the same physical server. They can also be migrated from one physical machine to another and scaled up or down according to the needs of the application. Figure 5 shows a schematic comparison between a traditional server and a virtualised server.

In essence, virtualisation technology enables a more dynamic capacity allocation in the data center. It also allows for consolidation of workloads (this term is explained in Section 2.4), making higher server utilisation possible (Barroso *et al.*, 2013d). The degree of virtualisation can be expressed in the average number of workloads per physical server, called *workload density,* and this is increasing in the overall server stock (Cisco, 2014).

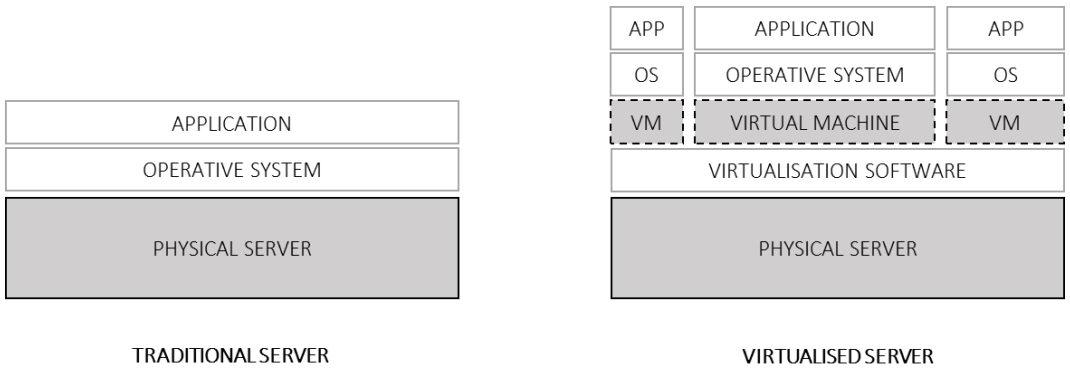

Figure 5: Traditional vs. virtualisation software architecture on a server.

An increasing degree of virtualisation in the server stock is highly related to the movement from server closets and rooms into large- and megascale data centers, as virtualisation software is more common in the latter case. The increase in workload density over the coming years will to a certain extent "absorb"

the increase in demand for data processing, since fewer servers can satisfy the same computing demand. This will matter to the energy demand of data center industry.

2.4 Workloads

A *workload* is a "computing job", caused by a request sent to a data center asking it to run an application or perform a digital service. Cisco (2014) defines a workload as follows:

> A server workload is defined as a virtual or physical set of computer resources, including storage, that are assigned to run a specific application or provide computing services for one to many users. A workload is a general measurement used to describe many different applications, from a small lightweight SaaS [software as a service] application to a large computational private cloud database application.

For example, writing "data center" in Wikipedia's search bar and clicking the search button will give rise to a workload in Wikipedia's data center, namely putting together the list of relevant articles and returning this result to my laptop.

As there are countless different services and applications, so there is an endless number of different workloads. Most of them can however be divided into two broad categories: batch workloads and interactive workloads, further described in the sections below. Many data centers handle a mix of interactive and batch workloads.

The workload request rate affects how busy a data center is, how many servers must be active and consequently how much power the data center needs. It is the variations in workload size and submission rate that bring about variations in data center power demand. The concept of workloads and their characteristics are therefore central to the demand-flexibility potential of data centers and consequently to remainder of this work.

2.4.1 Batch workloads

A batch workload is, as the name suggests, a relatively large workload in terms of computing demand. It is not unusual for this type of workload to have a running time of several hours or even days. Often, batch workloads also require large amounts of data to be sent between machines, e.g. extracts from a database on a storage server to be sent to the processing server. (Barroso *et al.*, 2013d)

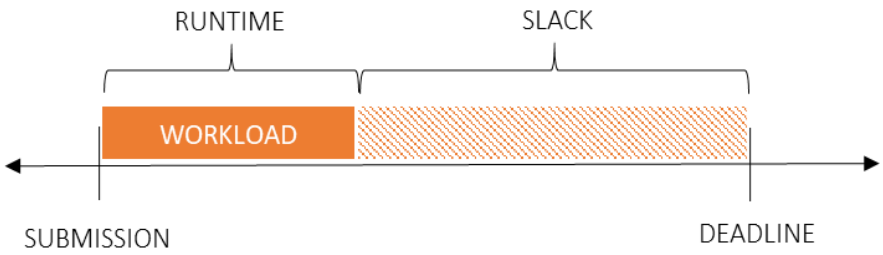

Figure 6: Illustration of a workload with slack.

The deadline and required resources (e.g. in CPU-hours) of batch workloads are often known at the point of submission to the data center. When the time between submission and deadline is longer than the required running time, which is often the case with batch workloads, it has a *slack*. This is illustrated in Figure 6. In other words, batch workloads often have a certain delay-tolerance, allowing for an intentional delay within the slack period without a violation of the deadline. (Liu *et al.*, 2012)

Financial analysis, image processing, system maintenance and scientific research (e.g. simulations) are examples of activities often producing batch workloads. (Liu *et al.*, 2012)

2.4.2 Interactive workloads

Interactive workloads demand a fast response. Any delay that occurs in the network connection between the client and the data center, or in the data center as a result of slow processing or queuing, may have negative impact for the user. (Liu *et al.*, 2012)

Submission times of interactive workloads are somewhat unpredictable. Though, since many of them are linked to human activity, interactive workloads often form a diurnal pattern. An example of this can be seen in Figure 7, where the CPU activity from a photo-sharing web service is plotted against a 7-day period. Another characteristic of interactive workloads is that, unlike for batch workloads, a relatively small amount of data is sent between machines. (Barroso *et al.*, 2013d)

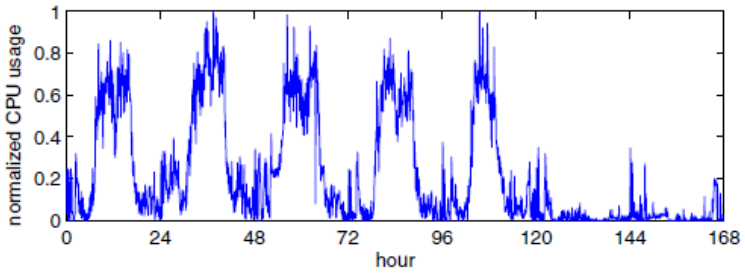

Figure 7: Workload trace from a photo-sharing web service. Adopted from (Liu et al., 2012).

Examples of interactive workloads are web applications or business transactional applications (Liu *et al.*, 2012). It is common that batch workloads are run to prepare for interactive workloads, for example by preparing a result list in a search engine. (Barroso *et al.*, 2013d)

2.5 Data center business

This section describes some of the commonly used terms used to classify different data centers and their services. The definitions are rarely strict and the different ways of categorising data centers overlap. Although the following subsections only briefly describe some business models and actor relationships in the industry, they highlight the heterogenous business environment and how it is common to have many actors and several stakeholders under the same roof in a data center. This can affect the willingness and ease to adapt to, for example, an energy conservation or demand-response scheme.

2.5.1 Cloud computing

A word often mentioned in a sentence along with data centers is *cloud computing*. Cloud computing refers to the service provided by many data centers, namely ubiquitous, on-demand access to storage or computing capacity from a shared pool of resources (NIST, 2011). Using a cloud service is hence an alternative to acquiring your own physical servers. A data center proving cloud computing is called *cloud data center* while non-cloud data centers are called *traditional* data centers.

The conditions of the service are defined in a service level agreement (SLA) which usually contains thresholds for e.g. latency, bandwidth and/or granted computing capacity. Latency is the delay or response time for the server when it is contacted by the client; bandwidth is the speed with which data can transferred in the network; and computing capacity is how much computing power the client will have access to.

Cloud service models

At what point in the data center architecture the cloud provider's service ends and the customer's own responsibility begins is described by different *service models*: Infrastructure as a service (IaaS), platform as a service (PaaS) or software as a service (SaaS), see Figure 8.

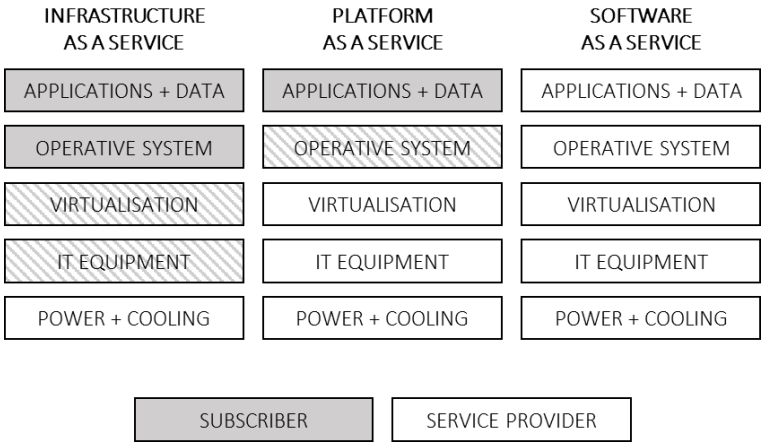

Figure 8: Division of responsibility between subscriber and service provider in different cloud service models.

2.5.2 In-house or colocation, on or off premises

Another distinction in the industry is whether the data centers are *in-house* or *colocation*. An in-house data center is purpose built by an organisation to serve the own needs and it is fully controlled by that organisation. These data centers tend to host a small number of applications but for a large number of users. The data centers of internet giants like Google, Facebook, Apple and eBay can be classified as in-house. A colocation data center is owned by an organisation that rents out computing capacity to many other organisations. It runs a large number of applications in a heterogenous software environment. (Länsstyrelsen Norrbotten, 2014)

Data centers can be *on-premises* or *off-premises* which simply refers to whether the data center facilities are found on the same location as the users, e.g. in or near the office space of an enterprise, or in a remote location. The previously mentioned trend of phasing out server closets and rooms and instead using cloud services hosted in large- or megascale data centers is generally also a movement from on-premises to off-premises.

3 ENERGY EFFICIENCY IN DATA CENTERS

In this chapter, the energy efficiency and performance indicators of data centers are described and their historical and present typical values are discussed.

3.1 Power Use Effectiveness (PUE)

The true efficiency of a data center should in theory be expressed in the amount of energy used to get one unit of useful work done. In practice however, this is an impossible benchmark due to the immense variety of types of workloads and applications.

Instead, the most widespread efficiency indicator in data centers is called *Power Use Effectiveness* (PUE) and measures the ratio between power used in the whole data center and that used in the IT-equipment. See Equation 1. It is easy to measure and calculate but is primarily an indicator of the efficiency of the supporting systems.

Equation 1: Power Use Effectiveness

$$PUE = \frac{Total\ data\ center\ power}{IT\text{-}equipment\ power}$$

PUE was made known by the organisation *The Green Grid* and has revealed some rather embarrassing inefficiencies in data center industry praxis. In 2006, 85% of data centers had a PUE of 3.0 or higher, meaning that for every 1 kW used in the servers, 2 kW were lost in somewhere else in the facility. There has been significant improvement though. Much of these losses were not due to physical constraints but due to poor design, primarily of the cooling system (Barroso *et al.*, 2013b). In 2014, the average reported PUE was down to 1.7 according to an Uptime Institute Survey and over half of respondents reported a PUE of below 1.5 (Uptime Institute, 2014). Figure 9 shows an average distribution of power consumption in a data center with a PUE of a little less than 2.

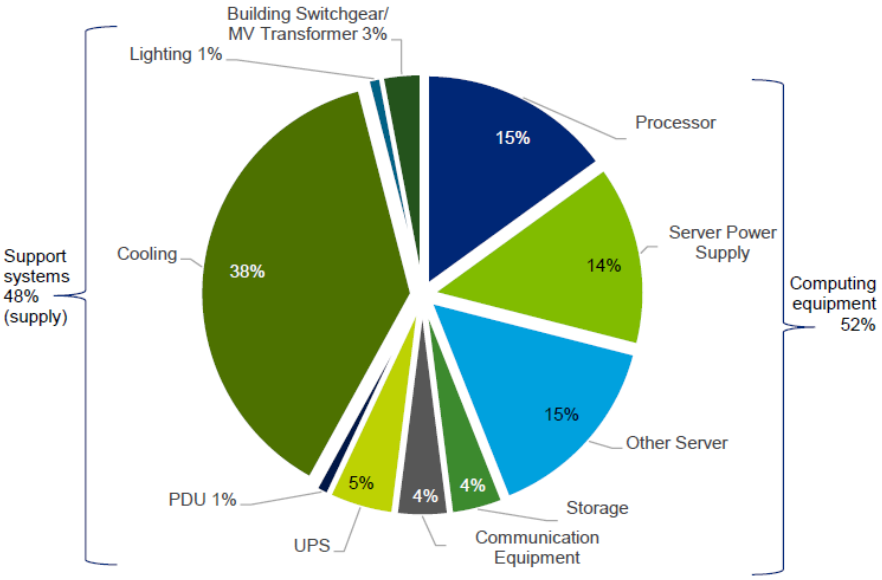

Figure 9: Average distribution of power consumption by component in a data center. Adopted from (European Commission, 2015).

Generally, larger data centers have a lower PUE than smaller ones (Commonwealth of Australia, 2014; European Commission, 2015). Therefore, a decrease in server closets and rooms and a subsequent

increase of their larger counterpart entails an overall PUE improvement. Some of the internet giants like Google and Facebook have as low as 1.1 (Facebook Sustainability, 2016; Google Environment, 2017).

3.2 Server efficiency

The overall efficiency of servers is increasing for every new product generation. Figure 10 shows the distribution, energy consumption and performance capability of servers from three different age categories. In 2015, the servers that were less than 5 years old used 35% of the total server energy but delivered 93% of the performance. As outdated servers are dismantled and replaced by new, the overall average server efficiency will increase. (European Commission, 2015)

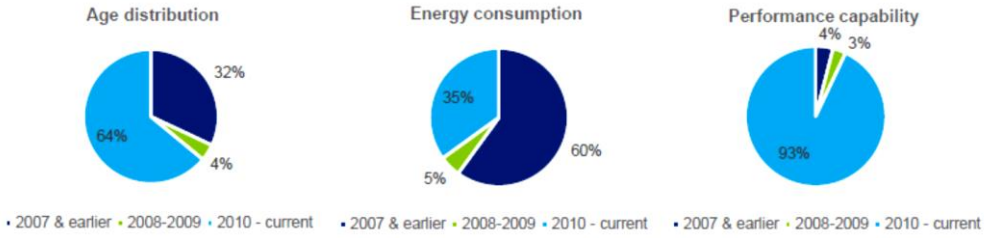

Figure 10: Age distribution, energy consumption and performance capability of servers, according to product generation. Adopted from (European Commission, 2015).

This improvement is inside the IT-equipment, independent of the surrounding cooling system. If the performance capability is kept constant, this server efficiency improvement would result in a smaller denominator in the PUE expression and consequently a higher value for PUE. In other words, more efficient servers give a worse energy performance according to the most wide-spread indicator. This exposes weakness of using only PUE as energy performance indicator for data centers.

3.2.1 Power proportionality

The momentary efficiency of a server is dependent on its level of utilisation and a term that relates the utilisation level and the power consumption of a server is *power proportionality*. In a perfectly power proportional server, the power draw (as a percentage of the maximum) is equal to the utilisation level throughout the utilisation spectrum. For example, a server with 100W maximum power draw would consume 10W at 10% utilisation, 50W at 50% utilisation, and so on.

In most servers though, the power consumption level is larger than the utilisation level, and especially so in the lower utilisation ranges. In other words, there is a lack of power proportionality. Meanwhile, average server utilisation in data centers is low; approximately 20% in the European Commission's estimate (2015). It is of course problematic that servers are not efficient in the utilisation span where they are likely to be working. It is also common for servers to draw over 50% of peak power at *idle state*, i.e. when turned on but at 0% utilisation.

Power proportionality in servers has nonetheless been improving too. Two diagrams from the *Standard Performance Evaluation Corporation* (SPEC) power benchmark database (SPEC, 2017) illustrate this in Figure 11. The left diagram shows the SPEC result for a Dell server from 2008 and the right one for a Dell server from 2017. The red bars show the performance to power ratio at 10% intervals of utilisation level. This is an indication of how much computing the server performs per watt it consumes. The blue lines show the power demand at each utilisation level.

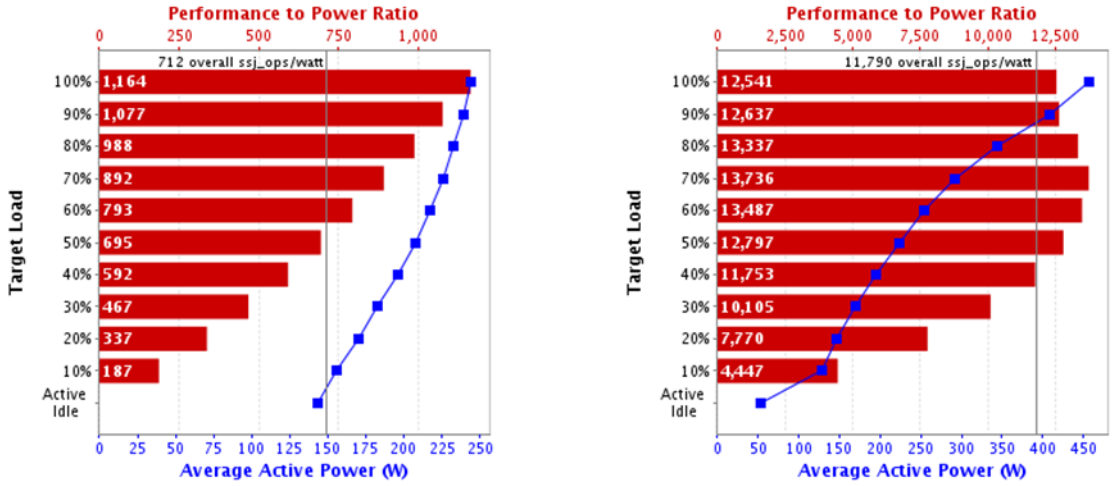

Figure 11: Performance to power ratio and average active power draw for (to the left) a Dell server from 2008 and (to the right) a Dell server from 2017. Source: (SPEC, 2017).

As you can expect from almost ten years of product development, the 2017 server has better performance per watt. What is perhaps more interesting in this context though, is how the blue power-to-utilisation curve has changed shape; the newer server's blue curve is bending towards the y-axis. This indicates that the server has been more energy-optimised for the lower utilisation spans. For example, at 20% utilisation level, the older server uses around 70% of the peak power. For the newer server this figure is down to 30%. A compilation of a larger number of SPEC power result sheets made by the European Commission (2015) proves that this change in design, making servers more efficient in lower utilization ranges, is a trend in the industry.

One might ask why we are seeing servers that are better designed to run at low utilisation, instead of a change in data center operations that increase the utilisation level in the servers. The problem is that the system architecture of many data center-based services complicates full server utilisation. In a search engine, for example, thousands of servers each manage one slice of the search index and are all working in parallel, but at low utilisation, to deliver the search result. Spreading data and functionality over a large number of servers also creates resilience and fail-tolerance that are important to and one of the major advantages of cloud services (Barroso *et al.*, 2013b). Improving the hardware at lower utilisation has therefore, it seems, been the preferred way to tackle this inefficiency.

3.2.2 Comatose servers

Comatose serves, also called zombie servers, are servers that are constantly idling. They are intended to allow for a quick reaction to an increase in workloads, but in reality they never get used and they are hence doing nothing but consuming power (Sverdlik, 2014). It is estimated that at least 20% of data center server are comatose (Heslin, 2018). The way to tackle this problem is to better understand the actual capacity needs of the data center applications and to decommission under-utilised servers.

3.3 Energy efficiency outlook

As the data processing industry grows, and its power demand with it, avoiding wasting energy will be increasingly important. Since the electricity costs are a substantial expense for data center operators, there is not just a green incentive but also an economic one to optimise the energy use.

From an energy conservation point of view, it makes good sense to consolidate servers and IT-services in large-scale cloud data centers. These can be made very efficient in terms of PUE and with good management the server utilisation, although it remains relatively low, can be raised when several organisations share resources. Shutting down comatose servers is, obviously, also a necessary measure to conserve data center energy. The currently large share of such servers indicates that data center

owners and operators do not have knowledge of their own system's needs and perhaps need better monitoring. With a raised level of specialisation in data center industry, which should come as a natural consequence of centralisation, there are good hopes of better such knowledge.

Power proportionality in servers and in the wider data center system is important not just to minimise wasted energy at lower utilisation levels. It is also a prerequisite for efficient resource management in the data center. If there is only minor difference between the power consumption between idle and active state, there is no reward in trying to efficiently dispatch server resources.

4 THE EUROPEAN DATA CENTER FLEET

In this chapter the composition of the data center fleet is assessed, with respect to scale, end-users and geographical distribution and the expected developments towards 2030 are identified.

4.1 Data center characteristics

The data center industry is heterogenous in many ways. In terms of scale, the IT-loads vary between a few kilowatts up to tens or even hundreds of megawatts. The scale distribution says something about how far along in the centralisation process this industry has come, which in turn also says something about the overall efficiency of the fleet. Furthermore, data centers are used across a wide range of sectors and for different end-user purposes. Different end-users have varying priorities and requirements for their data centers. Some therefore might be more commonly found in a certain type of data center or be more prune to certain changes than others. Finally, geographical distribution of data centers in Europe is related to the priorities made at the choice of location. Prioritising latency results in a data center in the proximity of the clients, prioritising efficient cooling should put the data center in a cold climate and prioritising green energy operations will often result in a site where such energy can be bought or produced at a reasonable cost. The following sections attempt to map the present European data center fleet according to these three aspects; scale distribution, end-user distribution and geographical hotspots.

4.2 Scale

A study from Germany will be used to find an estimate of the distribution according to data center scale. Although Germany is only one country, it is one of Europe's largest economies and home to many of its data centers[2]. It should therefore be indicative for Europe as a whole.

Case: Germany 2012

The study surveys the size distribution of German data centers (original source: Fichter and Hintemann, 2012; secondary source: European Commission, 2015). In the original source, data centers are categorised according to floor space, but the European Commission has converted this into average IT-load based on assumptions of the equipment's dimensions and rated power, see Table 2.

Table 2: Size ranges Germany

	Average IT-load (kW)
Server closet	2
Server room	7
Small data center	50
Medium data center	240
Large data center	2 500

Figure 12 shows the distribution of the number of data centers over the different data center scale categories in Germany 2012. It is clear from this chart that server closets and rooms are far more wide-spread than the larger data centers, in terms of number of sites.

[2] Germany was the largest contributor to EU total GDP in 2016, at 21.1%, according to Eurostat (ec.europa.eu/Eurostat). 12.6% of European data centers listed in datacentermap.com as of 2018-03-07 are located in Germany.

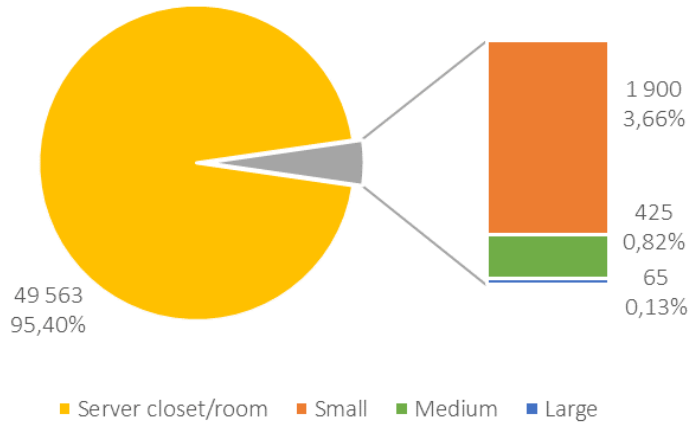

Figure 12: Number of servers and number of data centers according to data center scale in Germany 2012. Data source: (European Commission, 2015).

Figure 13 shows the different size categories' share of total data center energy consumption. This was calculated from the number of data centers in each category by the European Commission, using assumptions of energy efficiency. This figure shows that the relatively few 65 large data centers consume a comparable amount of energy as the almost 50 000 server closets and rooms together.

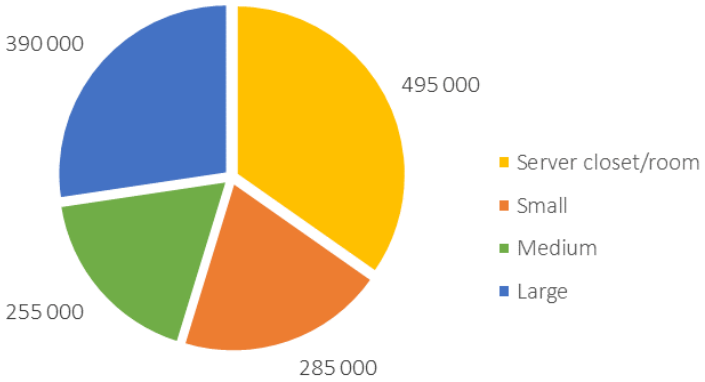

Figure 13: Distribution of energy consumption by data center size in Germany 2012 (piechart) and the number of servers in each category (figures next to chart). Data source: (European Commission, 2015).

Since 2012 when the data for this study was collected, there has been development in this rapidly changing industry. Already between 2012 and 2013 the number of large-scale data centers in Germany had increased from 65 to 70 (+ 7.6%) while server closet and rooms had gone down from 49 563 to 48 600 (- 1.9%) (Hintemann and Clausen, 2014). It is very likely that this trend has continued since. However, in lack of more recent data, the figures from 2012 may still provide an indication of the distribution today.

Cisco's hyperscale data centers

An industry report from Cisco can also be used to further shed some light on the prevalence of large- and megascale data centers in Europe. In Cisco's definition of hyperscale operators, 24 companies qualify. Although the definition of hyperscale is not based on the dimensions of the data center facilities, these data centers tend to be large- or megascale. The number of hyperscale data centers globally 2015

was 259. 42 of these, representing about 16% of the global fleet, were located in Europe – almost exclusively in Western Europe (Cisco, 2014).

At the end of 2017, Synergy Research Group (2017) reported that hyperscale data centers (by the same definition as Cisco) are up to 390 worldwide. Assuming, like Cisco, that the European share of these is now closer to 17%, the resulting number of hyperscale data centers in Europe is now around 66.

4.2.1 Outlook

Again, the ongoing trend concerning data center scale is that large- and megascale data centers are increasing in numbers, while server closets and rooms are decreasing. Cloud services, which are often hosted in large data centers, are growing in overall popularity, and are also replacing traditional solutions associated with local servers. According to the German study, the share of energy demand from data center in the MW-scale was already more than 25% and this share will grow in the years leading up to 2030.

4.3 End-user distribution

Two different sources are used to indicate the end-user distribution: Cisco's data on consumer and enterprise workload shares and a study from the Australian government on end use distribution.

Global consumer and enterprise workloads

Cisco has assessed the global number of workloads belonging to different end-users in 2015. The two categories – consumer and enterprise workloads – consist of a number of applications types. The consumer workloads are coming from: search; social networking; video sharing and streaming; and "other applications" including email, messaging, e-commerce, file sharing and more. Enterprise workloads are coming from: computing applications, which is mainly IaaS; collaboration, which includes email, conferencing, file sharing and social networking in a professional context; database, analytics and IoT, covering mainly database, big data and business intelligence applications; and "other enterprise applications", such as finance applications, service/systems/operations management, storage etc.

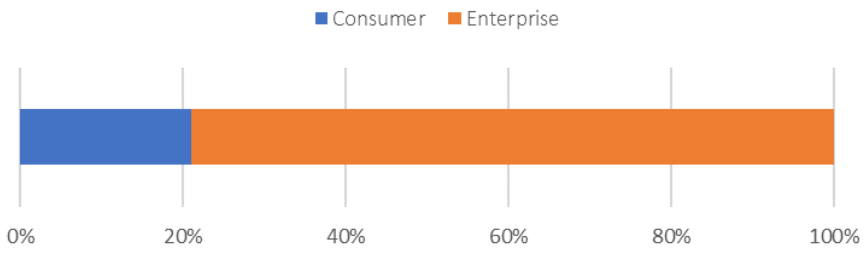

Figure 14: Consumer and enterprise share of global workloads in 2015. Data source: (Cisco, 2014).

In 2015, consumer workloads represented 21% and enterprise workloads 79% of global workloads, as illustrated in Figure 14. Note however that these figures are at a global scale. The fact that many global internet companies have many of their data centers in the US, but users all over world, means this consumer/enterprise ratio may not be representative for a region, e.g. Europe. These figures will instead serve as a starting point for discussing general changes in consumer/enterprise ratio.

Case: Australia

The shares of total data center energy consumption by different business sectors have been mapped for Australia and New Zeeland (Commonwealth of Australia, 2014). Though there might be differences in the economic landscape between different continents, the Australian breakdown still gives an indication of the situation in Europe, according to the European Commission (2015).

The Australian study divide the data center industry into four broad end-user categories: Government, finance and banking, telecommunications and media, and "other". Figure 15 shows the data center energy consumption by sector for data centers in Australia and New Zeeland in 2013. The energy consumption is calculated from original data on floor space.

This dataset unfortunately contains a mix of end-users and data center types and therefore also includes two service categories: "IT-services" and "colocation". IT-services are businesses that provide cloud services, to businesses and individuals. Colocation are businesses provide and operate the data center facility, support infrastructure and perhaps also the IT-equipment for several clients under one roof. Because customers of these services can be end-users belonging to other categories, there is an overlap between end-user and service provider shares. The shares of the four actual end-users could therefore be larger than represented here.

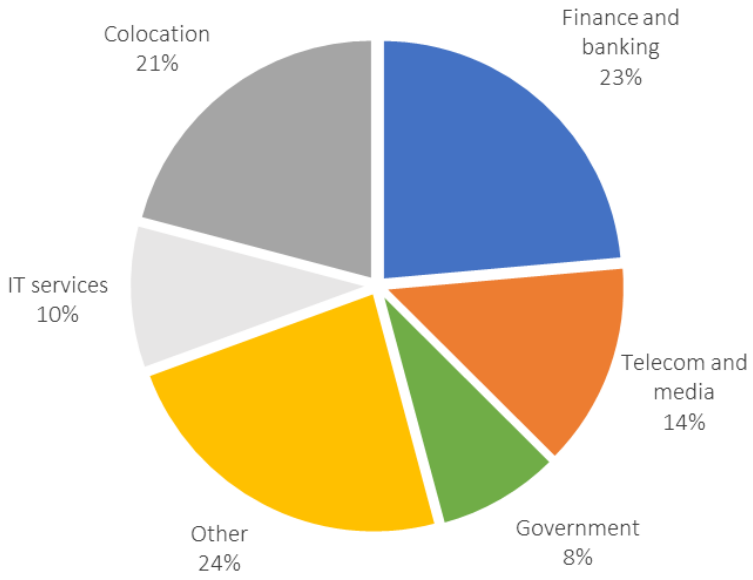

Figure 15: Energy consumption by sector in Australia and New Zeeland 2013.
Data source: (Commonwealth of Australia, 2014).

For most businesses and institutions, the data centers and their services have a supporting function; they provide tools, e.g. email, that are ancillary to the core business. Finance and banking and telecommunications are different in this aspect since the data center services are largely part of the core business, which may also be why these sectors have a larger share of the industry and of the consequent energy consumption.

There are some priorities and properties that are more commonly found in some business sectors than in others, in terms of their data centers and IT-services. Many departments of government agencies, as well as much of the finance and banking sector, have high requirements of security for their IT-systems. Furthermore, in finance and banking, the services hosted in the data centers include both processing large numbers of fast transactions and performing complex financial simulations and analysis. Their requirements therefore include both speed and high computing capacity for heavy data processing. Telecommunications data centers are often geographically distributed systems and need high resilience to carry out their role as the backbone of the internet. The media sector, which includes the fast-growing streaming services, have requirements of bandwidth, and it is common for data to be duplicated in several locations to ensure proximity to the users. "Others" is naturally a very diverse category. Here, it includes various smaller business sectors, manufacturing industry as well as healthcare and education. This results in heterogenous IT-service requirements throughout the category. An important note is that most server closets and rooms are likely to belong to this category. (Commonwealth of Australia, 2014)

Cisco's data in the previous section refers to number of workloads, while the Australian study shows energy use. Although the number of workloads is not necessarily proportional to energy use, the two are not unrelated as processing workloads undoubtedly requires energy. Comparing the studies, it seems that most of the Australian end-users – finance and banking, government and probably the better part of "others" – would belong to Cisco's enterprise share, while Cisco's consumer workloads would primarily come from telecommunications and media in the Australian definition. If the telecommunications and media share is completed with a few percentages that have been "lost" to the IT-services and colocation shares, it will not be far from Cisco's 21% consumer workloads. If for example the IT-services and colocation shares in the Australian pie chart are proportionally divided between the actual end-users, telecommunication and media reaches 20%. In conclusion, the two studies seem concurrent, even if they show slightly different things.

4.3.1 Outlook

Predicting end-users all the way to 2030 is very hard and just extending current growth rate would be a very uncertain method. For example, twelve years ago, Facebook was still just a social network for university students and few could predict that it would today have 1.4 billion daily users.

Cisco's projections of consumer and enterprise workload trends in the years 2015-2020 can perhaps also indicate the development in the following years. There is strong growth in the overall number of workloads, but in terms of consumer/enterprise ratio, consumer workloads are increasing their share of the total (from 21% to 28%), while enterprise share is dropping (from 79% to 72%). On the consumer side, social networking and video streaming are the fastest growing segments and will also be the two largest contributors to the total in 2020. For enterprises, compute and collaboration are the largest contributors to the total in 2020, but it is database, analytics and IoT that are growing the fastest.

Connecting these end-user distribution figures to the categorisation into batch and interactive workloads previously discussed is difficult to do with precision since the end-users probably all use a mix of batch and interactive workloads. However, very generally speaking, the social networking and video streaming on the consumer side are probably more interactive in their characteristics, while many enterprise workloads like compute, database, analytics and IoT are more batch-like.

4.4 Spatial distribution

There are some areas in Europe that are denser with data centers than others.

Self-listing

On the website *datacentermap.com*, colocation companies can list themselves as available for tenants. This database is not exhaustive but the largest publicly available listing of its kind. Of the currently 1488 colocation data centers listed in Europe, 1136 are in on the western half of the continent and the remaining 352 on the east. The five countries with the largest number of listed colocations providers are in descending order: UK, Germany, France, the Netherlands and Switzerland. (Datacentermap.com, 2018; Retrieved 4th April)

Synergy ranking

Another ranking, made by Synergy Research Group (2018), is of the world's top metropolitan areas in terms of revenue from colocation services. Four European cities are among these top 20: London, Frankfurt, Amsterdam and Paris. Synergy has also published a ranking of countries according to the number of hyperscale data center locations (Synergy Research Group, 2017). The US is very dominant with 44% of the global fleet but four European countries are among the following ten: UK at 6%, Germany at 5%, and Ireland and the Netherlands each at about 2%. Europe as a whole has 16-17% which means 2% are in other European countries.

It is perhaps not surprising that the data center hubs of Europe are in Germany, France, UK and their neighbouring countries since these areas are also the economic centers of Europe. This goes to show that a common strategy for choosing data center site is in the proximity of the business and/or customers, to lower latency and/or to have the facilities under own control.

4.4.1 Outlook

Data centers hubs are currently around important economic regions, like London, Paris and Frankfurt. Looking forward though, this is not necessarily where data center will be popping up in the next decade. Looking at data center risk index (Cushman & Wakefield, 2016), which ranks countries according to the lowest risk of operating data centers, with weighted factors of energy security, ease of doing business, labour costs and more. Data center risk index ranks the northern European countries higher. Iceland, Norway, Switzerland, Finland and Sweden are the top five worldwide. The next European country in that ranking is UK at 9 and Netherlands at 12.

*

This concludes the assessment of data center industry. We now move on to discussing flexibility.

5 DEMAND-RESPONSE FOR DATA CENTERS

In this chapter demand-response is introduced and two strategies for data centers to participate in it – workload scheduling and geographical load balancing – are described. The latter half of this chapter is a survey of green management strategies for data centers using workload scheduling and geographical load balancing, proposed and evaluated by the research community.

5.1 Renewable integration and demand-response

Before going into discussing data centers possible contributions in a more intermittent renewable power system, this section will briefly describe the renewable integration level and the concept of demand-response.

5.1.1 Renewable integration

By the year 2030, the European Union has set out to fulfil the following climate goals:

- Reducing the greenhouse gas emissions by 40% compared to 1990-levels
- Increasing the share of energy consumption originating from renewable sources to 27%
- Improving energy efficiency by 27% compared to projected levels

In a report from the European Commission's Joint Research Center, the composition of the European electricity mix in 2030 is projected for a scenario where these goals are met. In this so called EUCO27 scenario, the share of renewables in the gross generated electricity is projected to reach 47.3% 2030. Another scenario, EUCO3030, is a slightly more ambitious scenario, where the figures for renewables and energy efficiency goals are both raised from 27% to 30%. The resulting renewables share in the electricity mix is then projected to reach 54.2% in 2030.

Because wind and solar operate at variable power output, the future power system must have more installed capacity to get the same amount of energy, if compared to conventional thermal power technologies. For example, 1 GW installed nuclear provides 7.5 TWh electricity. To get the same amount of energy we need 2.5 GW installed wind or 5 GW installed solar (Åhman, 2016). In 2016, the installed renewable energy capacity in the European Union was 406 GW. For the scenarios EUCO27 and EU3030, the capacity in 2030 is projected to reach 652 GW and 712 GW respectively. The installed capacity and respective contributions from wind, solar, hydro and other renewables is shown in Figure 16.

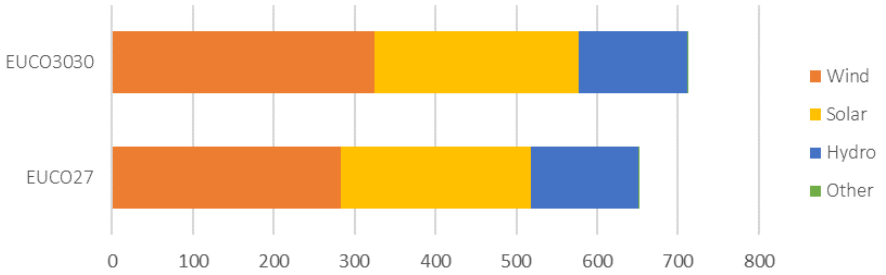

Figure 16: Installed capacity (GW) from renewable energy sources in the EU 2030, according to the EUCO3030 and EUCO27 scenarios. Data source: (Banja and Jégard, 2017).

The intermittent, non-dispatchable wind and solar is projected to make up between 500 and 600 GW installed capacity in 2030. As a consequence, the power output of the electrical power system will be much more variable than today – higher peaks and lower lows. More importantly, the output cannot be controlled. It is therefore necessary for some of the power demand to be flexible and able to adjust to the variable output of the power system.

5.1.2 Demand-response

Demand-response, or demand-side management, is an umbrella term for several actions that actors on the demand-side of the power system can take to change their electricity use. The objective of these actions can be to avoid peak electricity prices. It can however also be to secure the power system reliability, which is of interest to the grid operator and indirectly to the consumers. Typically, the actions consist of using less electricity at times of high demand and low supply, either by avoiding electricity use altogether, called *load-shedding,* or by shifting the use to a different time, called *load-shifting*. The latter is illustrated in Figure 17. In the case of an industry, load-shedding will most likely lead to a loss of production, for which the industry is likely to want some kind of compensation. With load-shifting, the total production is maintained but shifted in time. The incentive or compensation will only need to correspond to the inconveniences, if any, that are associated with this. (COWI/European Commission, 2016)

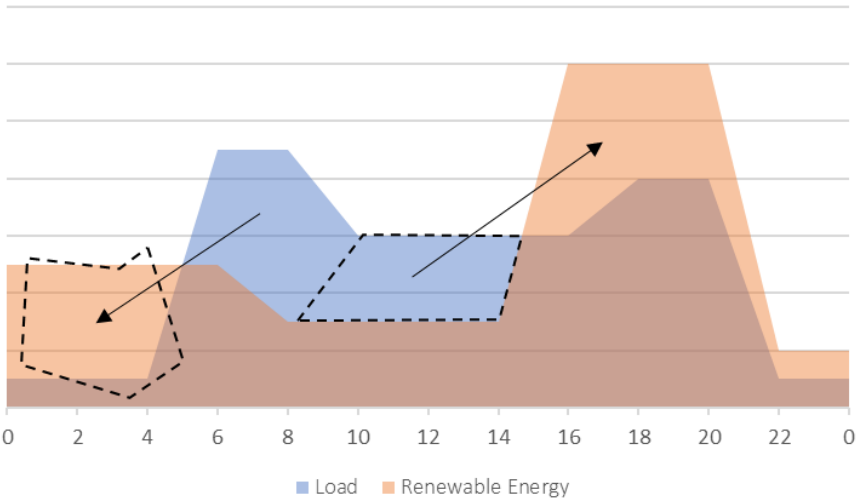

Figure 17: Illustration of load-shifting.

5.2 Demand-response strategies for data centers

Data centers are large electricity consumers with an inherently varying utilisation level. They are also highly connected and automated with components that are already monitored and responsive to control signals. These characteristics make them suitable candidates for demand-response, with minimal infrastructure investments (Krioukov *et al.*, 2011).

The variations in data center power demand ultimately depends on the processing of workloads. Aligning data center demand with energy supply in a high-renewable grid is therefore a matter of aligning the processing of workloads with renewable energy production. One strategy is to shift the workloads in time, which is here referred to as *workload scheduling*. Data center clusters also have a spatial dimension with the ability to direct workloads to different regions following local renewable energy production, here called *geographical load balancing*.

5.2.1 Workload scheduling

Workload scheduling builds on predictions of supply and demand i.e. the supply of renewable energy and the computing capacity demand of workloads. Using these predictions as input along with knowledge about the data center's server capacity, a scheduler (an algorithm) can create a workload time-sheet with the objective of aligning workloads and renewable energy supply. A prerequisite for renewable-following workload scheduling is that workloads have slack. Workload scheduling is hence a strategy suitable for data centers handling batch workloads. (Shuja *et al.*, 2016)

Predictions of renewable energy availability are based on weather data. Errors are inevitable, but predictions need often only be "good enough" to have a positive impact. Workload predictions can be based on historical data called *traces* from the application in question. Sometimes with batch workloads, the submission pattern is perfectly regular and hence known in advance. When submitted to the data center, the capacity demand and deadline of each batch workload is usually known. This information is an important enabler for workload scheduling.

From a grid point of view, workload scheduling data centers can provide load-shifting. The amount of energy used remains the same, but the power demand profile is changed. From the data center point of view this means that the amount of production (computing) will not be changed, but it will be shifted in time. If the shifting is done so that no deadlines are violated, it has no negative impact on the production/service.

5.2.2 Geographical load balancing

A data center operator who controls two or more facilities in different locations can also make use of this spatial dimension to match the data center activity level with energy supply. A geographical load-balancer builds on a mechanism similar to the workload scheduler but also makes decisions about *where* to run the workloads. By directing more workloads to the data centers with more renewable energy, and consequently less to others, data center operators can make the most of renewable energy in each location (Shuja *et al.*, 2016). Figure 18 schematically depicts a data center cluster with a geographical load balancer.

Geographical load balancing can be applied to both batch and interactive workloads. To avoid sending large amounts of data between remote locations though, a prerequisite for geographical load balancing with batch workloads is that the input data is replicated and available in more than one data center (Chen *et al.*, 2012). Because geographical load balancing does not necessarily involve more than a very marginal delay of the computing jobs, it can be applied to interactive workloads (Liu *et al.*, 2015).

When geographical load balancing is applied to interactive workloads, routing decisions can be made with a short time horizon or even in real-time. The small size of interactive workloads in terms of computing demand and input data allow them to be quickly routed between data centers, which makes planning ahead schemes unnecessary. (Toosi *et al.*, 2017)

From the grid point of view, geographical load balancing results in load-shedding. The load and energy demand that was originally destined for a data center in the local grid A but routed somewhere else, will not appear again in grid A. It will however instead appear in local grid B and increase the load and energy use there. Unlike for other types of industries who shed loads in a grid with high demand, the load-shedding does not lead to a loss of production for the data center operator, since the workload will still be processed somewhere else with only minimal impact on the service.

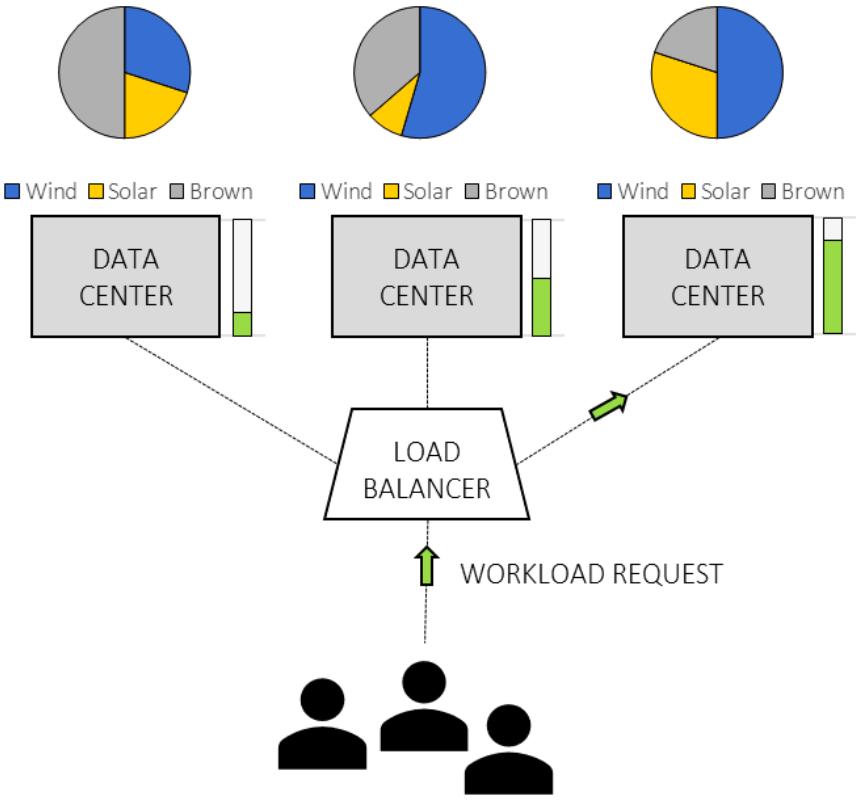

Figure 18: Schematic drawing of geographical load balancing in a data center cluster.

5.3 Research survey

There are several good reasons for data center operators to consider operating their data center in a flexible manner. Spatial and temporal workload management is a mean of avoiding congestion and queues in the data processing, which affects the quality of service. Since the electricity bill is a significant cost for operators, choosing to operate when rates are lower is a mean of cutting costs.

Adding the integration of renewables in the power mix to this equation has become increasingly interesting since many data centers invest in their own on-site renewables, for which they want a good return on investment. On a market where renewables are becoming cheaper and brown energy prices more volatile, following renewables can also cut costs and hedge against rising fossil fuel prices. Since attention has been drawn to the carbon footprint of the industry, it has also become a valuable sales point to run on green energy. These are some of the reasons why research efforts have been put into developing green workload management algorithms for data centers.

The remainder of this chapter is a survey of seven such research works. Through this survey, the qualitative potential of data centers as flexible loads will be assessed. These seven studies have been selected for the survey on the following grounds:

- They all model and evaluate workload scheduling or geographical load balancing in data centers, with the objective of increasing the use of renewable energy and decreasing brown energy demand.
- They benchmark this against a conventional workload management scheme.
- Between them, they represent a variety of scenarios and assumptions, in terms of workload characteristics (batch and/or interactive), renewable energy source (wind and/or solar) and location of the facilities.

5.3.1 Study A: Matching renewable energy supply and demand in green data centers

A study from Goiri *et al.* (2014) evaluates the potential for workload scheduling in a data center running batch workloads. This data center is located in New Jersey, USA and is partly powered by on-site solar panels. The modelling is done at a small scale with a cluster of 16 servers with common case peak-power of 2.3kW and the solar resource is scaled to cover this at peak power output.

The load scheduling algorithm proposed in this study plans the workloads with the objective of making the best use of the solar power without violating the deadlines. If the available solar power is insufficient to complete workloads on time, the servers draw power from the grid, which is assumed to be dominated by brown energy. The algorithm also considers the variable cost of grid power and seeks to minimise it.

A simulation is run to evaluate the scheduler in this setup. The solar energy data in the simulation is calculated from widely available weather data and the workloads are based on traces from real applications. One workload trace is from scientific application where workloads are submitted on a regular basis. The other trace is from Facebook Hadoop and consists of data processing in a distributed storage system, with random submission times. The former workloads each run on one server, while the latter contains jobs that are run parallelly on many servers.

The results are a 19-21% increase of green energy use for the Hadoop workload and a 11-117% increase for the scientific workloads. This is when compared to a conventional scheduler that seeks to finish the workloads as fast as possible. The wide span in the results is due to different weather conditions in the example weeks that were studied. For an average week the result for the scientific workload is 45%. Furthermore, if the algorithm ignores brown energy costs, the green energy is increased by 47% for an average week.

5.3.2 Study B: Integrating renewable energy using data center analytics systems: Challenges and opportunities

A study by Krioukov *et al.* (2011) evaluates workload scheduling on a server cluster partly running on wind power and that handles batch workloads. The server cluster in the model consist of 576 servers located in California, USA. In the model, the wind power comes from several sites in the proximity of the data center but are directly attached, as opposed to grid connected. The ratio between the wind resource and the data center power draw is altered in the simulations but found optimal at around 1:1.

The scheduling algorithm proposed in this study has a pragmatic approach. For each scheduling interval it first runs workloads that require immediate action. It then checks for availability of wind power and schedules additional workloads in the order of shortest remaining slack first. This does not lead to an optimal solution, but it is deemed good enough.

The workloads used in the evaluation of the algorithm are of scientific nature. The trace comes from a language processing application which runs jobs parallelly on several servers. The slacks for these workloads are generally 40-80 minutes. The wind energy data is based on measurements from real wind farms in major wind regions in California.

The supply-following scheduling algorithm is benchmarked against an algorithm that runs all workloads immediately upon available server capacity. Using the green workload scheduling algorithm results in roughly 60% of the data center power being drawn from the wind resource. For the conventional scheduler this share is 35%. This equals 38% reduction in grid power draw, assuming that the total power consumption is unchanged.

This algorithm's strategy can lead to bottlenecks causing workloads to exceed their deadline, but it is rare (< 0.5% of the workloads) and the delays are only small fractions of the execution times. Its advantage is that even if workload and wind traces contain inaccuracies, chances are good that the workload slacks can be fully exploited for harnessing wind power.

5.3.3 Study C: Renewable and cooling aware workload management for sustainable data centers

Liu et al.'s (2012) study of green workload scheduling considers a data center with on-site solar panels and incorporates the fact that outdoor temperature affects the cooling performance. The model data center has 500 servers corresponding to 100kW, a solar array of 130kW and cooling system with outdoor air cooling and a chiller. It handles a mix of interactive and batch workloads.

The scheduling algorithm executes all interactive workloads immediately upon submission while the batch workloads are subject to scheduling. If the solar power is insufficient, the data center runs on grid power. The objective of the scheduler is to minimise grid power use and optimally use the own solar power, without violating deadlines. At midnight every day, the scheduler creates an optimal running plan for the batch workloads during the following 24h, based on the predicted energy availability, server and cooling power demand and with respect to the interference of anticipated interactive workloads.

The interactive workloads in the simulation are modelled from an international web service trace. The batch workloads are submitted regularly at noon and midnight and represent 1.5 times the computing resource demand of the interactive workloads. The solar energy supply and cooling demand data are collected from a real data center.

Results show that the green scheduler reduces grid power draw by 39% compared to running the workloads immediately or at an even pace throughout the 24h. Worth noting though, is that operations with the green scheduler draws slightly more energy in total (+2%).

5.3.4 Study D: Green-aware workload scheduling in geographically distributed data centers

The study by Chen, He and Tang (2012) applies a load scheduling policy to a geographically distributed data center cluster. The default model in this study is of two data centers in California, USA and in Hong Kong. The model is also extended with two more data centers in Virginia, USA and in Spain. Each data center has 480 servers. Associated with each data center is a solar farm with a peak power output equivalent to the data center peak power demand.

The algorithm first identifies workloads that require immediate action and dispatches them to the data center with the minimum brown energy consumption, before dispatching the remaining workloads in order of shortest remaining slack. The decision about where to migrate a workload to is based on the green energy availability and the energy usage of servers and cooling, but also with respect to data transfer needs. If the solar power supply is insufficient, data centers use brown grid energy.

The workloads in the evaluation of the algorithm are of scientific nature, based on a trace from a two thousand-node cluster at Los Alamos Lab from the year 2000. Of the 6799 workloads in this trace, 8% have a running time of less than 10 seconds and 10% have a running time of over 10 hours, but the workloads can be divided into smaller jobs and tasks. The solar energy input data come from real solar panel traces during a random week in May.

The green scheduler is benchmarked against RoundRobin (RR), a scheduler oblivious to green energy and cooling demand. With two data centers, the green scheduling algorithm uses 53% green energy, while RR uses 46%. This can be expressed as a 14% reduction in brown energy usage. When the number of data centers is increased to four, the resulting brown energy reduction is 40%. The study also concludes that the time zone difference between the data center locations has large influence on the result in this setup.

5.3.5 Study E: Greening geographical load-balancing

Liu *et al.* (2015) approach geographical load balancing as a trade-off between energy cost and performance cost (increased delay).

The modelled cluster consists of 14 data centers located at the centre of ten American states, and it handles interactive workloads. The data center operators buy dynamically priced electricity from the utility. The grid power is a mix of brown energy and renewable energy from equal parts solar and wind and price signals indicate the availability of renewables. High share of renewables results in a lower price creating an incentive to migrate workloads.

The approach in this study is that an increased service delay, caused by workload migration, imposes a cost. This cost is balanced against the energy cost when workloads are routed among data centers in different locations. The delay is assumed to be proportional to the distance between the data centers.

The workload trace used in this simulation is from Hotmail, a web-based email service. This trace is used to build a workload request source in each state, shifted according to time zone and scaled according to the internet connected population. The renewable energy production is constructed from weather data.

This study shows that, provided there is an economic incentive to follow the renewables, i.e. that green energy is cheaper than brown energy, this approach can create a shift in data center power draw between different regions, following the availability of renewable energy. The figure below illustrates this. The left diagram shows the generation of renewable energy on the US East and West coast respectively. The middle diagram shows the distribution of active servers in the two regions, which follows the renewables under a pricing regime with a strong incentive to do so. The middle diagram should be compared to the right diagram, which shows the same server distribution but under static pricing, i.e. no incentive to follow renewables.

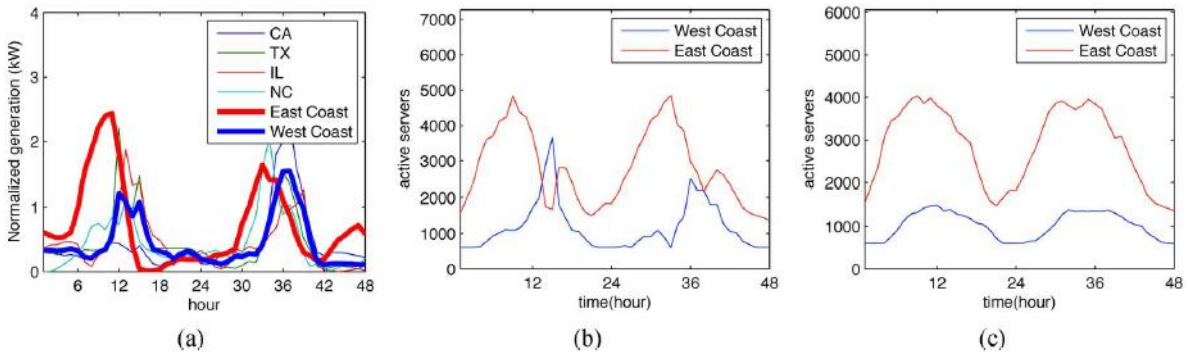

Figure 19: Results from Liu et al. 2015. a) Renewable energy availability in different regions, b) Server utilisation in different regions under an economic incentive to follow the renewables, c) Server utilisation in different regions without an economic incentive to follow the renewables. Adopted from (Liu et al., 2015).

5.3.6 Study F: GreenWare: Greening cloud-scale data centers to maximise the use of renewable energy

The data center cluster in Zhang, Wang and Wang's (2011) model is distributed over four American states, including one location in Hawaii, and handles interactive workloads. The data centers draw their power from the local utility grids, each of which has a certain amount of wind and solar power in the energy mix. The green energy is more expensive than brown energy.

To deal with the objective of minimising brown energy usage without causing unacceptably large power bills, the algorithm proposed in this study maximises the renewable energy while working within a pre-set budget. The quality of service (response time) is always maintained.

The workloads are modelled from a Wikipedia trace and the availability of renewable energy is calculated from assumptions of installed wind and solar capacity and weather data.

When benchmarked against an algorithm that seeks only to minimise cost, which in this setting means choosing brown energy only, the green algorithm achieves a 42% decrease in brown energy at a 52% cost

increase. When no real restriction is imposed by the monthly budget, the data centers are powered by on average 60% renewable energy. The results in this study are of course highly affected by assumptions of budget and energy pricing.

5.3.7 Study G: Renewable-aware geographical load balancing of web applications for sustainable data centers

The system model in Toosi et al.'s (2017) study is of a data center cluster with servers in three cities in France. Each site has a wind turbine and solar panels, scaled so that their combined average production covers the servers' average demand. The data centers host an interactive web-application.

The load balancing algorithm seeks to distribute the requests among data centers so that, if possible, all requests are handled using renewable energy without affecting the quality of service. Should the renewable energy not be enough to handle all requests, the surplus is executed with non-renewable energy in the data center with the cheapest electricity price.

The workload trace used is again from Wikipedia and the renewable energy production in the is calculated from historical weather data.

Toosi et al.'s load balancer is benchmarked against Round Robin which spreads workloads evenly among data centers in the cluster. The green load balancer uses 17% less non-renewable energy with no significant degradation of service quality.

5.3.8 Summary of survey

These studies exemplify both workload scheduling and geographical load balancing, by letting an algorithm make workload management decisions informed by renewable availability. A generic decision logic of the workload management algorithms can be found in Appendix A.

In studies A-C, workload scheduling is applied to batch workloads in a single data center. In all three cases, delaying of workloads to match with renewable energy production is only permitted if it can be done without violating workload deadlines. If not, workloads are still launched on brown/grid energy. Study D is an example of geographical load balancing of batch workloads, while study E-G are geographical load balancing of interactive workloads.

The studies are formulated from the perspective of the data center operator, whose concerns are primarily managing costs and environmental impact (carbon emissions) associated with their data center. The results from the flexible operations of the data centers are therefore expressed as a percentage reduction of brown (grid) energy, compared to a conventional scheduler. The studies and their results are summarised in Table 3.

Table 3: Overview of research studies included in the survey and summary of their results.

	Authors	Workloads	RE source	Single /Cluster	Location	Reduction of brown/grid energy
A	Goiri et al. 2014	Batch	Solar	Single	New Jersey, USA	19-21% for Hadoop Avg. 45% for scientific
B	Krioukov et al. 2011	Batch	Wind	Single	California, USA	38%
C	Liu et al. 2012	Batch and interactive	Solar	Single	California, USA	39%
D	Chen, He & Tang 2012	Batch	Solar	Cluster of 2-4	Global distribution	14% for two DC:s 40% for four DC:s
E	Liu et al. 2014	Interactive	Wind and solar	Cluster of 14	USA	Not applicable
F	Zhang, Wang & Wang 2011	Interactive	Wind and solar	Cluster of 4	USA	42% under proposed budget-cap
G	Toosi et al. 2017	Interactive	Wind and solar	Cluster of 3	France	17%

Power profiles

In the interest of assessing demand-response potential, it is not quite satisfactory to know only the share of green energy in the overall power mix. To get a sense of the change in power profile, we will therefore look closer at some of the studies that have shown how their algorithm has affected the power use the data center during the course of the simulation.

Liu et al. 2012 concerns workload scheduling in a data center handling a mix of batch and interactive workloads. The benchmark case, to the left in Figure 20, runs batch workloads at an even pace throughout the day, in which case this consumes roughly a constant 30% of the total IT-capacity. The green scheduler gathers batch workloads in the sunny hours of the day when renewable energy is available and at night when outdoor temperatures are low and interference with interactive workloads is minimal. This is shown to the right in Figure 20. During these hours the batch workloads instead consume 60-70% of the total IT-capacity, which along with the interactive workloads reaches 100% of IT-capacity during the day.

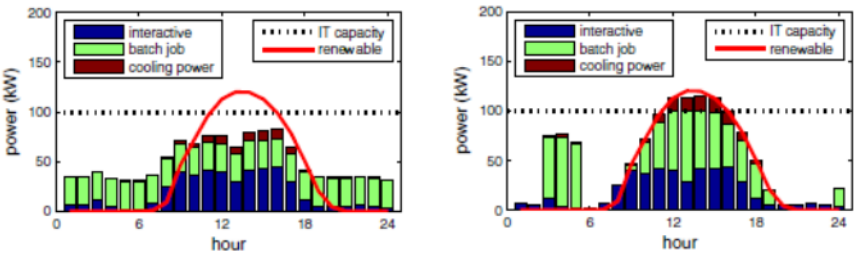

Figure 20: Results from Liu et al. 2012. Left: Running batch workloads flat throughout the 24h. Right: Workload scheduling to match renewable energy availability. Adopted from (Liu et al. 2012).

The power use profiles from Toosi et al. 2017, whose study concerns geographical load balancing are also interesting in this context. The figures cannot be reproduced in this work due to copyright restrictions but can be found in the original publication. In the renewable-oblivious case that is used for benchmarking in this study, the power use of the data centers in all three locations is ramping up and down between roughly 0.5 kW during the night and 0.9 kW during the day in a quite regular fashion. Under a geographical load balancing scheme, the variations in the activity levels increase. The night activity level goes down to 0.2kW in the solar powered data centers and workloads are instead routed the data center with wind power production during the night, resulting in a 1.0-1.5kW power use there. The daytime power use goes up to a little over 1.0kW in the solar powered locations.

5.4 Load-shifting opportunities

This survey goes to show that it is indeed possible for data centers to operate in a supply-following way, both in terms of timing and geography, with little or no impact on the quality of service. In these studies, the objectives are formulated from a data center operators point of view. Although the data center operator objective of increasing the share of green energy in the consumed power mix is not directly equal to a grid-operator objective of balancing availability and demand, the two are related. The load-flexibility that has been demonstrated in the studies should be possible to leverage in demand-response schemes. In other words, there is a qualitative potential for data centers to engage in demand-response.

6 FLEXIBLE DATA CENTER POWER 2030

In this chapter a rough approximation is made of the power available for demand-response from data centers in 2030, based on different energy scenarios. The results are then put into perspective with projections of the power system 2030.

6.1 Data center energy scenarios for 2030

The basis of the calculations is two data center energy scenarios found in previous studies. One represents a conservative, low projection of European data center use, and one represents a fast-growth, high scenario.

6.1.1 European Commission (EC-Low)

The first data center energy scenario to be used in this analysis comes from the European Commission's Ecodesign Preparatory Study (European Commission, 2015). The study contains projections of the annual energy demand from data centers in the EU-28 from 2015 to 2030 for a number of different scenarios.

A server stock projection is the basis for all scenarios. The stock projection is based on sales figures, life-time assumptions and expected demand growth, and the results are checked for consistency against an alternative approach based on workload projections. The workload growth is based on figures for 2013-2017 published by Cisco in 2013 and extended to 2030 by the European Commission by assuming a maintained compound annual growth rate (CAGR). The resulting workload projection is depicted in Figure 21.

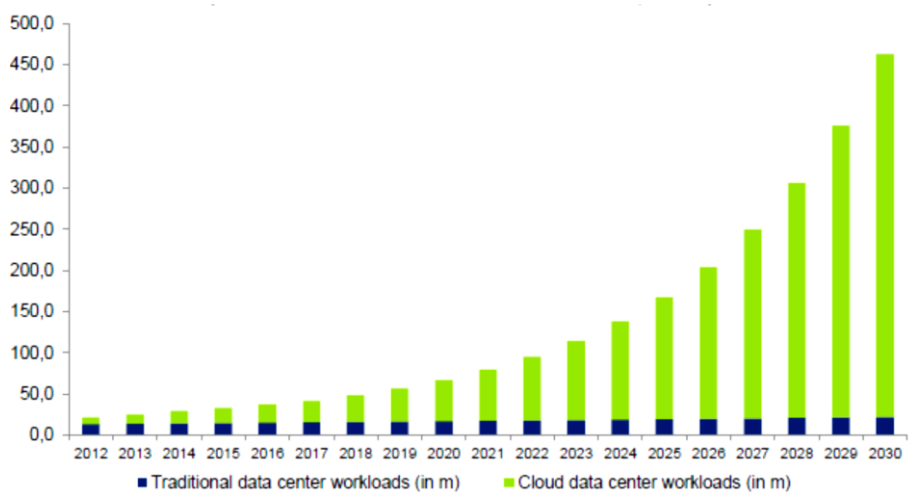

Figure 21: Workloads for Europe 2012-2030. Adopted from (European Commission, 2015).

Between 2014 and 2030, the EU-28 server stock is expected to grow from around 9 million to around 13 million units. It is assumed that chip-performance and virtualisation level in the servers will increase by the same rate it did between 2010 and 2012, which is enough to absorb much of the drastic increase in workloads projected. In fact, the improving server performance in the coming years is projected to cause server stocks to decrease until 2020, before picking up growth again towards 2030.

Based in this server stock development, different scenarios are then created assuming different PUE developments and policy initiatives, see Figure 22. The "ambitious business as usual" scenario (BAU_Amb) will be chosen and adopted in this work. In the Ecodesign study, the BAU_Amb assumes no direct policy action, but a market-driven drop in average PUE from 2.0 in 2015 to 1.5 in 2030. Though

called “ambitious”, this is still a relatively conservative efficiency improvement for the supporting infrastructure.

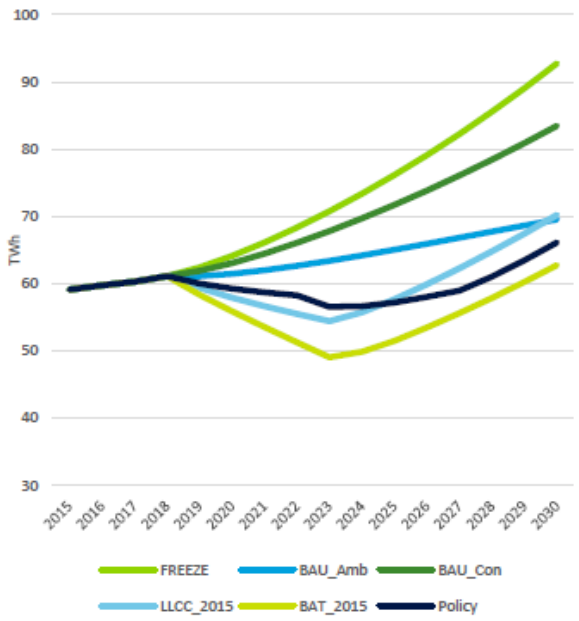

Figure 22: Scenarios from the Ecodesign study. Adopted from (European Commission, 2015).

The method used in the Ecodesign study, a base case analysis, implies that in 2015, EU data centers consumed 60 TWh of electricity. The authors of the report however comment that based on overall findings in their study, the real figure for 2015 is rather 78 TWh. This discrepancy will be taken into account when interpreting the results of BAU_Amb and creating a low-growth scenario for 2030 to be used in this work.

The base case analysis for BAU_Amb suggests a data center energy consumption of 69.5 TWh in 2030. With the adjustment for a projection undershoot of the same proportions as in the 2015 case described above, the energy consumed by data centers in 2030 amounts to 90.4 TWh. This figure will represent the first energy scenario in this study and will be referred to as the *EC-Low*.

Note that the conclusion of the Ecodesign study is a relatively modest increase in energy demand from data centers, from 78 TWh in 2015 to 90.4 TWh in 2030. Though explicitly taking into account a drastic increase in demand for IT-services (expressed in workloads), the study expects much of this to be absorbed by significant performance gains in the servers, which are “due to increased chip performance (Moore’s law) and utilisation rates through virtualisation”. Therefore, even with a conservative PUE, the result is only a slight increase in energy demand from data centers.

The Ecodesign study could classify as a bottom-up projection, where the authors have tried to identify individual factors, drivers and causes (server sales and life cycles, efficiency potentials etc.) and add them together towards a result.

6.1.2 Andrea and Edler (AE-High)

Andrea and Edlers projection is based on growth in global IP traffic and data center electricity usage in 2010-2011, and then assumes an “annual electricity efficiency improvement”. Both server performance and energy efficiency improvements of data center infrastructure are merged into this figure. For their “expected” scenario, this improvement is 10% until 2022 and thereafter 5%. This scenario along with the “best” and “worst” case are depicted in Figure 23.

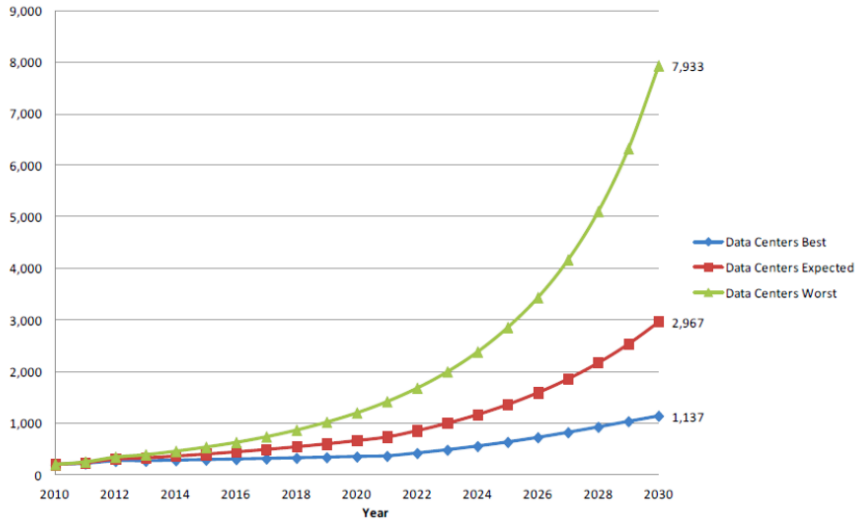

Figure 23: Estimates of global data center electricity usage (in TWh/yr). Adopted from (Andrae and Edler, 2015).

According to Andrea and Edler projection, the global data center electricity demand to be expected in 2030 is 2,967 TWh. Around 20% of the world's data centers are currently located in Europe. Provided that this share remains the same and that it directly reflects the energy demand, the European share of the global total should in this scenario amount to 593 TWh per year in 2030. This will serve as a high-growth scenario, *AE-High*, in this work. For comparison, the 2015 figure by the same logic is around 80 TWh for Europe, i.e. similar to the European Commission's 2015 estimate.

This study is better described as a top-down calculation, where the previous growth in data center electricity consumption is the main basis of the projection.

6.1.3 Compromise scenario (CK-Middle)

The AE-High projection for 2030's data center energy demand is more than six times higher than the that of EC-Low. Notably, the two scenarios encompass two slightly different geographical areas; EC-Low is limited to EU member states while AE-High roughly accounts for the European continent. This should however only have marginal impact.

The two studies differ in their bottom-up versus top-down approach. More precisely, a major difference between AE-High and EC-Low is the assumed relationship between:

a) the growth in IP traffic and workloads, which reflects the growing demand for the services provided by data centers; and
b) the performance and efficiency gains in servers and ancillary systems

While EC-Low assumes that a) will be largely absorbed by b), AE-High assumes a) will outpace b).

It seems likely that both scenarios above represent the extremes in the range of possible data center energy demand 2030. It is optimistic from an energy conservation point of view to assume that server performance can absorb almost all added demand, like the Ecodesign study does. The referenced Moore's law, which describes how chip-performance has been improving exponentially, is slowing down as manufacturers are approaching a physical limit (Simonite, 2016). It also seems like an overshoot to assume, like in Andrea and Edler projection, that data center energy demand will grow by the same exponential rate for decades. This would imply a doubling of the energy demand, from around 300 TWh to almost 600 TWh in the last 4 years leading up to 2030.

A reasonable energy demand for data centers in 2030 is therefore likely to be somewhere between the roughly 90 TWh and 600 TWh that the scenarios suggest, and probably closer to the lower value. A

middle ground scenario called CK-Middle at 200 TWh is therefore added to the following analysis. The annual energy demand in the three scenarios are summarised in the third column in Table 4.

6.2 Calculation of available power

Making a "correct" calculation of demand-response potential of data centers in 2030 would require a detailed power system model of Europe, complete with locations of renewable sources, data centers and other demand-side agents, transmission lines and more. Furthermore, this model would have to be run for a full year on hourly basis. Such a study would require research efforts far beyond the scope of this master thesis, and since much of the input would be based on assumptions, it would still not be sure to result in reliable numbers.

This study does not attempt to model the future power system. The aim of the calculation in this work is merely to estimate of the order of magnitude of the data center flexible power reserve, with the purpose of demonstrating the possible potential of this industry's participation in demand-response in the future.

6.2.1 Average power demand

The power demand is synonym to the load that data centers pose to the grid. The power demand of the data center fleet is hence the basis of how much load that could be shifted or shedded. The annual energy demand will therefore be converted to terms of power demand.

First, an average power demand (P_{avg}) is deducted from the annual energy demand in the energy scenarios. Since data centers have a near-constant uptime, P_{avg} is easily calculated by dividing the annual energy demand E_{ann} by the 8760 h in a year.

$$P_{avg} = \frac{E_{ann}}{8760}$$

P_{avg} is the mean value of all European data center's added power demand over the course of the year 2030. It can also be described as the total power demand if all European data centers ran at a constant rate throughout the year.

Applying this calculation to the annual energy demands from the different scenarios yields an average power demand of 10.3 GW for EC-Low, 67.7 GW for AE-High and 22.8 GW for CK-Middle for 2030. See Table 4.

Table 4: Annual energy demand (E_{ann}) and average power demand (P_{avg}) in the three scenarios.

Scenario	E_{ann} 2015	P_{avg} 2015	E_{ann} 2030	P_{avg} 2030
EC-Low	78 TWh	8.9 GW	90.4 TWh	10.3 GW
AE-High	80 TWh	9.1 GW	593 TWh	67.7 GW
CK-Middle	-	-	200 TWh	22.8 GW

6.2.2 Utilisation and power consumption pattern

The momentary power demand of a data center depends on the current level of utilisation. The utilisation level profile of a data center varies largely depending on the design, software architecture and applications in question. In this step of the calculation it is therefore necessary to make some high-level generalisations, based on average utilisation levels.

The average server utilisation in a data center is currently around 20%. More IT-services are expected move into cloud data centers in coming years and since these consolidate workloads and share resources, the average server utilisation level is likely to rise. For 2030, let us therefore assume that the average server utilisation level is 30% for European data centers.

Let us further assume that this average is in fact made up by a 16h-period of 45% utilisation and an 8h-period of 0% utilisation per 24h period, as depicted in Figure 24 below. This is of course a simplification but reflects the fact that activity levels for many IT-services are linked to human rhythm of day (European Commission, 2015). This approximation is also supported by several of the workload traces encountered in the survey in the previous chapter.

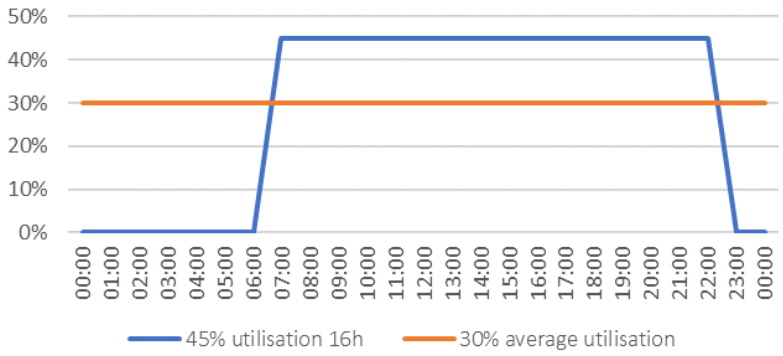

Figure 24: Simplified approximation of the server utilisation level over a 24h period.

The power use of the server and of the supporting equipment is often disproportionately high compared to server utilisation level, especially for the low ends of the range. Most importantly, data centers still consume a substantial amount of power when servers are idling, i.e. when they are responsive but at 0% utilisation. With better design of servers and data centers the gap between utilisation and power consumption level will shrink, but design constraints prevent it from closing completely.

Recent performance-to-power data from real servers, such as the right diagram in Figure 11 previously in this work, shows a roughly linear relationship between the server utilisation and power consumption as a share of peak power. We further assume here that the power demand of the cooling system and other ancillary equipment changes continuously and proportionally with the server utilisation level. This again is a simplification since in reality cooling systems are likely to be upgraded in a number of discrete steps.

With this simplification we get an approximation of the relationship between the server utilisation and the data center power draw. The relationship between server utilisation level and data center power draw is approximated as linear, with a starting point in 20% power consumption level for idle state which reflects an anticipated decrease of idle power draw. See Figure 25.

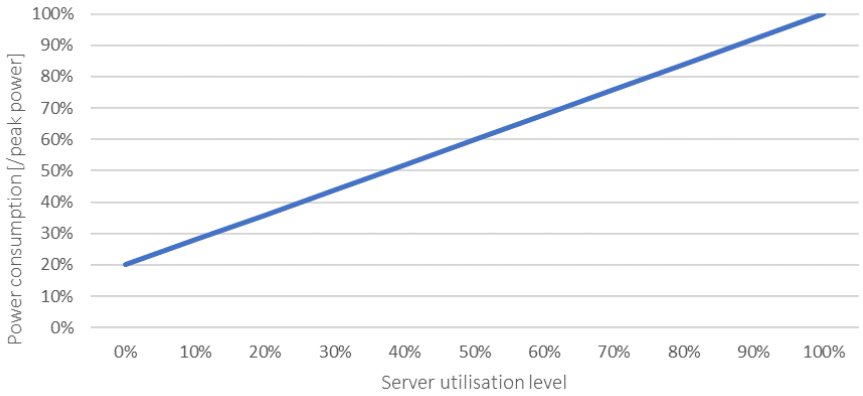

Figure 25: Approximation of the relationship between server utilisation and power consumption for an average data center in 2030.

6.2.3 Power range

Let us hypothesise now that the average European data center power demand P_{avg} in the section above is the mean value from a European data center fleet that is following the utilisation pattern in Figure 24, i.e. switching between 0% and 45% server utilisation levels. Note that this assumption implies that servers are always turned on and responsive, which is common practice in data centers.

We now want to find out what the two power demand values, $P_{0\%}$ and $P_{45\%}$, are. The relationship between P_{avg}, $P_{0\%}$ and $P_{45\%}$ are described in Equation 2.

Equation 2

$$24 * P_{avg} = 8 * P_{0\%} + 16 * P_{45\%}$$

The power levels for the two different server utilisation levels, $p_{0\%}$ = 20% and $p_{45\%}$ = 56%, come from the linear relationship in Figure 25. These relate the two power consumption values in Equation 3.

Equation 3

$$\frac{P_{0\%}}{p_{0\%}} = \frac{P_{45\%}}{p_{45\%}}$$

Solving this equation system formed by Equation 2 and 3 allows us to express $P_{0\%}$ and $P_{45\%}$ in P_{avg}.

$$P_{0\%} = 0.45 P_{avg}$$

$$P_{45\%} = 1.27 P_{avg}$$

$P_{0\%}$ is the European data center power use in a moment when all data centers are simultaneously idling. In practice, this event will never occur, but this power level gives us the theoretical lower boundary for data center flexible power (without involving switching on and off servers). The $P_{45\%}$ similarly represents the power use in a moment when all data centers are running at 45% utilisation, which is here assumed to be a normal average server utilisation level for a data center during the busier part of a day.

Using the expression for $P_{0\%}$ and the relationship between utilisation levels 0% and 100% we can also determine the power use for 100% utilisation. This is, even more so than for $P_{0\%}$, not a realistic event, but a theoretical boundary. It represents a moment when all data centers are simultaneously running at their peak capacity.

$$P_{100\%} = 2.27 P_{avg}$$

The difference between 0% and 100%, is hence the theoretical amount of power for which European data centers can be flexible in their demand. The difference between 0% and 45%, which is also presented, is the power difference between the approximated daytime and night time average working points.

In the conservative energy scenario EC-Low, the total power range between idling and full performance of European data center is estimated to 18.7 GW. Assuming that all data centers have a high activity (day) and low activity (night) period, the total amount of power between these utilisation levels is estimated to 8.4 GW. In the high growth energy scenario AE-High, the estimated full power range reaches 120.0 GW, and the day/night utilisation power range 55.4 GW. Finally, in the CK-Middle scenario the full range is estimated to 41.5 GW and the day/night range to 19.7 GW. These results are summarised in Table 5.

Table 5: Average power demand (P_{avg}); idle, "busy" and peak power demand ($P_{0\%}$, $P_{45\%}$ and $P_{100\%}$); the day/night power range ($P_{45\text{-}0}$); and the full idle/peak power range ($P_{100\text{-}0}$) for European data centers in the three scenarios.

Scenario	P_{avg}	$P_{0\%}$	$P_{45\%}$	$P_{100\%}$	$P_{45\text{-}0}$	$P_{100\text{-}0}$
EC-Low	10.3 GW	4.7 GW	13.1 GW	23.4 GW	8.4 GW	18.7 GW
AE-High	67.7 GW	30.8 GW	86.2 GW	153.8 GW	55.4 GW	120.0 GW
CK-Middle	22.8 GW	10.3 GW	30.0 GW	51.8 GW	19.7 GW	41.5 GW

The calculation has been made for all three scenarios, as a form of sensitivity analysis with respect to the assumed annual energy demand. As explained in section 6.1.3, the third scenario CK-Middle is considered the most reasonable of the three. Therefore, the following calculation involving workload characteristics is only carried out for CK-Middle.

6.2.4 Accounting for workload types

For data center's actual ability to be flexible power consumers, some constraints are set with respect to their core purpose – IT-services. As previously explained, a high-level categorisation can be made between computing jobs in batch workloads and interactive workloads. Most data centers deal with a mix of these two, and how much of each type are found in each data center depends on the applications and IT-services in question.

In the study by Liu et al. (2012), which was part of the survey, a data center with a mix of batch and interactive workloads is modelled. In this set-up it is assumed that the computing demand of batch workloads is 1.5 times that of the interactive workloads. Although this will not be true for every data center, and it hard to know if it will be representative of the average in over a decade from now, we will apply this assumption to the CK-Middle scenario.

Table 6 shows the potential amount of flexible power associated with each category of workload, for the two flexible power ranges $P_{45\text{-}0}$ and $P_{100\text{-}0}$ in the CK-Middle energy scenario, based on the 1:1.5 ratio between interactive and batch workloads.

Table 6: Flexible power associated with batch workloads and interactive workloads respectively, in the CK-Middle energy scenario.

	P_{batch}	$P_{interactive}$
$P_{45\text{-}0}$	11.8 GW	7.88 GW
$P_{100\text{-}0}$	24.9 GW	16.6 GW

Since batch workloads often have a certain delay-tolerance, there is an opportunity to schedule them and their associated power demand. The delay sensitivity of batch workloads is entirely depending on the service and application from which they arise. In the survey in an earlier chapter of this work though, slacks were commonly in the hour-range. P_{batch} hence shows an approximation of how much data center power could be available for hour-day-range scheduling, if all European data centers in 2030 participated.

The interactive workloads cannot be scheduled because of high delay-sensitivity. The power associated with the interactive workloads, $P_{interactive}$, are instead the theoretical amount of power that could be geographically shifted by routing requests to other data centers. This however requires that a large number of data centers belong to a cluster and are able to participate.

6.3 Summary of results

The result of the calculation for the CK-Middle scenario, which is considered as a reasonable middle way between the two different data center energy projections, are the following:

- The European data center energy use in 2030 is assumed to reach 200 TWh.

- The total installed load of data centers is then estimated to roughly 50 GW, but the average power use is 20 GW.
- As an indication of the order of magnitude, data centers could represent 20-40 GW of power available for demand-response.

An impact assessment study of demand response in the EU (COWI/European Commission, 2016) can be used to put the results of this work into perspective. In the report, the peak load in the EU in 2030 is estimated to 568 GW. The maximum theoretical demand-response potential is estimated to 28% of this, corresponding to 160 GW. Data centers were not listed among the appliances and processes whose demand-response potential was considered. Part of data centers' power use might however qualify as the listed process "industrial cooling".

The estimated potential of data center demand-response hence corresponds to up to 25% of that of all considered appliances and processes. The most promising contributor to demand-response according to the impact assessment study is "electric vehicles/batteries", at 60 GW, followed by commercial buildings at a little less than 20 GW. Data center potential is in a comparable order of magnitude to this.

7 DISCUSSION

Here the findings in this study are summarised and discussed and conclusions are presented about data center industry's future role in the energy system and its demand-response potential.

7.1 Data centers in the future energy system

The increasingly important role of data and digital infrastructure in our lifestyle and economy should be evident to everyone. Watching films, navigating cities, paying bills and attending meetings are only a few examples of activities that have partly or entirely become internet-based over the course of the past years. As a result, it has become normal and quite necessary for people to always carry a smartphone in their pocket – a device that only a few decades ago would be considered a palm-size supercomputer with cameras and sensors worthy of a space expedition. With Internet of Things, big data, artificial intelligence, blockchain technology and all other digital innovations in the pipeline, this development is not about to slow down. Data centers are a vital part of the infrastructure that supports this digital way of life. This industry is a backbone in a digitalised society and its importance is only going to increase. It is understandable why it will be a large and growing contributor to our total electricity demand for the foreseeable future and probably far beyond 2030.

Accurately and reliably projecting the future energy demand of data center industry is however difficult, if not impossible. The two energy projections referred to in this study – from the European Commission and Andrea and Edler respectively – demonstrated the large uncertainties, with results that differ by almost a full order of magnitude. It is not necessary to know the exact figure though, to realise it will be significant to the energy system. Consider for example that the total electricity demand for all industrial sectors (excluding the energy sector) in the European Union currently amounts to about 1000 TWh annualy (Eurostat, 2018). Even the most conservative data center energy projection for 2030 amounts to 9% of this.

The growing energy demand of data centers has been on the agenda in recent years, in the interest of energy conservation. Yet this emerging industry is rarely considered from an energy system perspective. Many publications, roadmaps and plans about the future energy system fail to recognise and acknowledge this potentially important piece.

As of today, a large share of data processing still takes place in small installations – server closets and rooms – that are distributed and hidden inside office buildings. This makes their electricity demand spread out and "invisible". Perhaps this is one reason data processing in not yet viewed as an industrial activity. By 2030 though, many of these small in-house IT-solutions will have been replaced by cloud services hosted in large-scale cloud data centers in the megawatt-range. The electricity demand will hereby be more centralised and more closely resemble that of a present-day industrial process.

This work has focused on data centers integration in the electrical power system, and specifically on demand-response potential in scheduling and routing of workloads. This is only one of many energy system aspects worth assessing. The large amounts of waste heat from data center cooling could be integrated in district heating systems (examples of this already exist e.g. in Stockholm) or low-temperature heat applications, possibly within agriculture. Furthermore, data center back-up power systems could be based on fuel-cell technology instead of diesel generators and be integrated in the hydrogen economy.

It is likely that most of the large- and megascale data centers of 2030 have not been built yet. This means that the opportunity exists now to be proactive and avoid lock-in effects. It also allows the new industry to evolve with a new kind of energy system. One aspect to consider is of course to build energy efficient data centers and establish best practices, both for the design and operation. Initiatives towards this have already been taken. Data centers can earn a variety of voluntary certifications and enterprise servers are about to become part of the Ecodesign directive, which applies to the European market.

Another aspect to consider is *where* to build new data centers. With an increased adoption of cloud services, a company's servers do not have to be in the same place or even the same country as the users. This means data centers can be built in more energy optimal sites, e.g. in cold climates where cooling is efficient and/or in a location where they can run on renewable energy with minor transmission losses. This also reduces the need to compete about space in regions where land is scarce and perhaps better used for other purposes.

The conclusions drawn about data center industry are:

- Data centers are a new energy intensive industry. They should be considered, assessed and understood from a strategic energy system perspective.
- Most large-scale data centers of 2030 have not been built yet. Policy measures should be used to assure that the new industries are built according to best practice.

7.2 Demand-response opportunities

To find out if and how data centers can modify their power consumption, this work contained a survey of previous research on green workload management strategies for data centers running partly on intermittent renewable energy. It was evident from this survey that data centers can be operated in a supply-following manner to achieve objectives interesting to the data center operator. Typically, this is to minimise energy cost without compromising the quality of service. It would be possible to use the same workload management mechanisms to achieve objectives from a grid-operator's point of view, such as balancing a high-renewable power grid through demand-response.

Scheduling batch workloads according to renewable energy availability is perhaps the most easily achievable demand-response strategy for data centers. Because many batch workloads have slack (delay-tolerance) it is possible to schedule them within a certain period without negative impact on the service. As this strategy only involves one data center location it can be adopted by single data center operators. The time horizon for this load-shifting opportunity will depend on the typical slack of the workloads in question. In the survey, most batch workloads had a slack in the hour-range. This means workload scheduling could substitute or lessen the need for hour-day energy storage, like batteries. Notably this puts workload scheduling data centers in the same demand-response market as electric vehicle fleet, which might lead to competition.

For data center operators who manage two or more data centers in different regions there is also the possibility of geographical load balancing. This has the particularly interesting feature of allowing for load-shedding in a strained local grid, without a loss of production for the data center operator. This could mean good business for data centers capable of load-shedding through geographical load balancing at the grid operators request, since this service would be compensated. Routing workloads from data centers in regions with a power shortage to data centers in regions with power abundance, instead of transmitting the electrical energy in the opposite direction, could cut transmission losses and/or ease congestion in the transmission grid. Geographical load balancing might however impose a slight delay. If it is applied to interactive workloads it might therefore negatively affect the quality of service. For a cloud provider, it may therefore be necessary to change praxis in service level agreements and lower the latency requirements to allow for geographical load balancing. Perhaps not all customers will suffer from a few additional milliseconds and instead value a sustainable operation of the data center that host their application.

There are many other actors and appliances in society who could potentially be engaged in demand-response – electric vehicles, home appliances, thermal inertia in buildings. An advantage that data centers might have over many of them is that one central actor can control an entire MW-range power reserve. That means there is only one actor to convince, educate or whatever type of communication is needed to get them on-board for the demand-response. For some data centers though, there might be several stakeholders with split interests at different levels in the data center ownership structure, for example in a colocation data center, which could make an introduction of demand-response difficult.

However, the number of people involved are still likely to be few. Compare for example with electric vehicles or home appliances where each car/home owner must be committed.

In conclusion:

- Through workload scheduling data centers can load-shift on an hour-day horizon. This could substitute batteries.
- Through geographical load balancing data centers can load-shed on local grids, which could mitigate grid congestion.

7.3 Power potential

The deducted estimate of this study is that data center total available load for flexibility is in the order of magnitude 20-40 GW for Europe in total. This puts data center demand-response potential right between the other promising demand-response providers electric vehicles and commercial buildings.

Part of this potential will be in the form of load-shifting through scheduling of batch workloads. Based on the assumption that batch workloads represent 60% of the total computing demand, this part of the potential is in the range of 12-24 GW. In the calculations leading to this result, it was assumed that servers are never switched off. This is common praxis in many data centers as it allows for quick response to an increase in workload requests. However, when applying workload scheduling the computing demand of workloads will be known for a foreseeable future which should open up for planned shut-downs of servers that are not needed. This in turn allows for even deeper cuts in the power demand from the data center during periods of low energy availability. When incorporating server on- and off-switching, the demand-response potential from workload scheduling might be higher than estimated.

The other part of the estimated potential will be in the form of geographical load balancing. Assuming this part is 40%, it amounts to 8-16 GW. For all this potential to be realised, it would be necessary for data centers hosting web services and other fast-response applications to be part of a cluster, where a centralised workload management system can route workloads between several regions. If only part of the data centers can be operated in this way, this potential can only be partly utilised. On the other hand, the data centers that belong to global actors are likely to be part of a global cluster. The larger span of those clusters would probably increase the demand-response potential of those particular data centers, but it is difficult to say by how much.

Forecasting the future is usually associated with large uncertainties and the calculation carried out in this work is no exception. Having a rough estimate is however better than having no idea at all. The fact that is places data centers in a similar range as other promising demand-response providers indicates that the deducted figure in fact could be quite accurate.

The potential available for demand-response as well as other grid services could probably also be increased by involving the back-up power systems (batteries and generators) present in practically every free-standing data center.

The conclusion is therefore:

- The available power for load-shifting is estimated to be in the order of magnitude of 20-40 GW.

7.4 Concluding note

Not seizing the demand-response potential of the emerging data center industry would be a waste. ENTSO-E (2016) estimates that 150 billion euros in pan-European significant storage and transmission investments is required only to accomplish the 2030 goal of 27% renewables. The infrastructure investments need to engage data centers in demand-response are merely some code and perhaps some metering. It is simply a very resource efficient way of dealing with the challenges of a high-renewable power system.

REFERENCES

Andrae, A. and Edler, T. (2015) 'On Global Electricity Usage of Communication Technology: Trends to 2030', *Challenges*, 6(1), pp. 117–157.

Banja, M. and Jégard, M. (2017) *Renewable technologies in the EU electricity sector: trends and projections*. Luxembourg: Publications Office of the European Union (publication nr. JRC109254). Available at: https://iet.jrc.ec.europa.eu/remea/publications/all-publications.

Barroso, L. A., Clidaras, J. and Hölzle, U. (2013a) 'Data Center Basics', in Hill, M. D. (ed.) *The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool Publishers, pp. 47–65.

Barroso, L. A., Clidaras, J. and Hölzle, U. (2013b) 'Energy and Power Efficiency', in Hill, M. D. (ed.) *The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool Publishers, pp. 67–88.

Barroso, L. A., Clidaras, J. and Hölzle, U. (2013c) 'Introduction', in Hill, M. D. (ed.) *The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool Publishers, pp. 1–13.

Barroso, L. A., Clidaras, J. and Hölzle, U. (2013d) 'Workloads and Software Infrastructure', in Hill, M. D. (ed.) *The Datacenter as a Computer - An Introduction to the Design of Warehouse-Scale Machines*. Morgan & Claypool Publishers, pp. 15–32.

Chen, C., He, B. and Tang, X. (2012) 'Green-Aware Workload Scheduling in Geographically Distributed Data Centers', in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. Taipei, Taiwan: IEEE, pp. 82–89.

Cisco (2014), *Cisco Global Cloud Index : Forecast and Methodology, 2015–2020*. Cisco Public (White paper).

Commonwealth of Australia (2014), *Energy Efficiency Policy Options for Australian and New Zealand Data Centres*. Department of Industry (Report).

Corcoran, P. and Andrae, A. (2013), *Emerging trends in electricity consumption for consumer ICT*. (Working paper). Available at: https://wireless.kth.se/wp-content/uploads/sites/19/2014/08/Emerging-Trends-in-Electricity-Consumption-for-Consumer-ICT.pdf.

COWI/European Commission (2016), *Impact assessment study on downstream flexibility, price flexibility, demand response & smart metering*. DG Energy (request nr. ENER/B3/2015-641). Available at: https://ec.europa.eu/energy/en/studies/impact-assessment-study-downstream-flexibility-price-flexibility-demand-response-smart.

Cushman & Wakefield (2016) *Data Centre Risk Index 2016*. Available at: http://www.cushmanwakefield.com/en/research-and-insight/2016/data-centre-risk-index-2016.

Datacentermap.com (2018) *Colocation Data Centers*. Available at: http://www.datacentermap.com/datacenters.html (Accessed: 10 April 2018).

ENTSO-E (2016) *Ten-year network development plan 2016*. Brussels: ENTSO-E (Report).

European Commission (2015) E*codesign Preparatory Study on Enterprise Servers and Data Equipment*. Luxembourg: Publications Office of the European Union. Available at: https://publications.europa.eu/en/publication-detail/-/publication/6ec8bbe6-b8f7-11e5-8d3c-01aa75ed71a1/language-en.

European Union (2018) *Climate Action*. Available at: https://europa.eu/european-union/topics/climate-action_en (Accessed: 2 January 2018).

Eurostat (2018) Electricity consumption by industry, transport activities and households/services (GWH). Available at: http://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=ten00094&language=en (Accessed: 8 May 2018).

Facebook Sustainability (2016) *Our Footprint*. Available at: https://sustainability.fb.com/our-footprint/ (Accessed: 30 April 2018).

Fichter, K. and Hintemann, R. (2012) *Energieverbrauch und Energiekosten von Servern und Rechenzentren in Deutschland. Aktuelle Trends und Einsparpotenziale bis 2015*. Borderstep Institut. Available at: https://www.borderstep.de/publikation/hintemann-r-fichter-k-2012-energieverbrauch-und-energiekosten-von-servern-und-rechenzentren-deutschland-aktuelle-trends-und-einsparpotenziale-bis-2015-kurzstudie-rechenzentren/.

Goiri, Í., Haque, M. E., Le, K., Beauchea, R., Nguyen, T. D., Guitart, J., Torres, J. and Bianchini, R. (2014) '*Matching renewable energy supply and*

demand in green datacenters', Ad Hoc Networks, 25(February), pp. 520–534.

Google Environment (2017) *Google 2017 Environment Web Report*. Available at: https://environment.google/environmental-report/ (Accessed: 30 April 2018).

Heslin, K. (2018) *Decommissioning as a Discipline: Server Roundup Winners Share Success*. Uptime Institute. Available at: https://journal.uptimeinstitute.com/decommissioning-discipline-server-roundup-winners-share-success/ (Accessed: 1 March 2018).

Hintemann, R. and Clausen, J. (2014) *Rechenzentren in Deutschland: Eine Studie zur Darstellung der wirtschaftlichen Bedeutung und der Wettbewerbssituation*. Borderstep Institut. Available at: https://www.borderstep.de/publikation/hintemann-r-clausen-j-2014-rechenzentren-deutschland-eine-studie-zur-darstellung-der-wirtschaftlichen-bedeutung-und-wettbewerbssituation/.

IPCC (2014) *Summary for Policymakers, Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Edited by O. Edenhofer, R. Pichs-Madruga, Sokona Y., E. Farahani, S. Kadner, Seyboth K., A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlömer, C. von Stechow, T. Zwickel, and J. C. Minx. Cambridge, New York: Cambridge University Press.

IRENA (2018) *Renewable Power Generation Costs in 2017*. Abu Dhabi: International Renewable Energy Agency.

Koomey, J. (2011) 'Growth in Data Center Electricity use 2005 to 2010', *Analytics Press.*, pp. 1–24.

Krioukov, A., Goebely, C., Alspaugh, S., Chen, Y., Culler, D. and Katz, R. (2011) 'Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities', *IEEE Data Eng. Bull.*, 34(1), pp. 3–11.

Liu, Z., Chen, Y., Bash, C., Wierman, A., Gmach, D., Wang, Z., Marwah, M. and Hyser, C. (2012) 'Renewable and Cooling Aware Workload Management for Sustainable Data Centers', in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*.

Liu, Z., Lin, M., Wierman, A., Low, S., Andrew, L. L. H. and Member, S. (2015) 'Greening Geographical Load Balancing', *IEEE/ACM Transactions on Networking*, 23(2), pp. 657–671.

Länsstyrelsen Norrbotten (2011), Ansökan om tillstånd enligt miljöbalken: Dnr 551-2464-11.

Länsstyrelsen Norrbotten (2014), *Strategi för att skapa en världsledande teknikregion i Norrbotten för klimatsmarta effektiva datacenter*. Luleå: Länstyrelsen i Norrbottens län (Report).

Metayer, M., Breyer, C. and Fell, H.-J. (2015) 'The projections for the future and quality in the past of the World Energy Outlook for solar PV and other renewable energy technologies', in *31st EU PVSEC*.

NIST (2011), *The NIST Definition of Cloud Computing*. National Institute of Standards and Technology (Special Publication 800-145). Available at: http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

OECD/IEA (2016) World Energy Outlook 2016. Paris: IEA Publications.

Shuja, J., Gani, A., Shamshirband, S., Ahmad, R. W. and Bilal, K. (2016) 'Sustainable Cloud Data Centers: A survey of enabling techniques and technologies', *Renewable and Sustainable Energy Reviews*. Elsevier, 62, pp. 195–214.

Simonite, T. (2016) Moore's Law Is Dead. Now What?, *MIT Technology Review*. Available at: https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/ (Accessed: 8 May 2018).

SPEC (2017) SPECpower_ssj2008 *Results, Standard Performance Evaluation Corporation*. Available at: https://www.spec.org/power_ssj2008/results/ (Accessed: 4 January 2018).

Sverdlik, Y. (2014) Survey: Industry Average Data Center PUE Stays Nearly Flat Over Four Years, *Data Center Knowledge*. Available at: http://www.datacenterknowledge.com/archives/2014/06/02/survey-industry-average-data-center-pue-stays-nearly-flat-four-years (Accessed: 14 February 2018).

Sverdlik, Y. (2017) Research: There are Now Close to 400 Hyper-Scale Data Centers in the World, *Data Center Knowledge*. Available at: http://www.datacenterknowledge.com/cloud/research-there-are-now-close-400-hyper-scale-data-centers-world (Accessed: 22 December 2017).

Synergy Research Group (2017) *Hyperscale Data Center Count Approaches the 400 Mark; US Still Dominates*. Available at: https://www.srgresearch.com/articles/hyperscale-

data-center-count-approaches-400-mark-us-still-dominates (Accessed: 14 February 2018).

Synergy Research Group (2018) *Just 20 Metros Generate 59% of Worldwide Colocation Revenues. Available* at: https://www.srgresearch.com/articles/just-20-metros-generate-59-worldwide-colocation-revenues (Accessed: 14 February 2018).

Toosi, A. N., Qu, C., de Assunção, M. D. and Buyya, R. (2017) 'Renewable-aware geographical load balancing of web applications for sustainable data centers'*, Journal of Network and Computer Applications*. Elsevier Ltd, 83(January), pp. 155–168.

Uptime Institute (2014) *2014 Data Center Industry Survey*. Available at: https://journal.uptimeinstitute.com/2014-data-center-industry-survey/ (Accessed: 4 January 2017).

Zhang, Y., Wang, Y. and Wang, X. (2011) 'GreenWare: Greening cloud-scale data centers to maximize the use of renewable energy', *Lecture Notes in Computer Science*, 7049, pp. 143–164.

Åhman, M. (2016) *Elmarknadens omvandling - Reglering, vägval och drivkrafter för elsystemets utveckling till 2050*. Lund: Lund University, Department of Technology and Society, Environmental and Energy Systems Studies (Report nr 96).

APPENDIX

Appendix A

Generic decision tree of green workload management algorithms in chapter 5.3.