1. Introduction

Data centers, as increasingly huge energy consumers, can assume a new role in the future energy system: instead of demanding power and energy as needed, they might adapt their power profile to the requirements of the power grid through more or less automated communication and trading channels. This concept is called demand response, the temporary adaptation of power demand to economic incentives like varying prices or contracts with the electricity power grid service provider. Fostering this concept with data centers as participants is the main goal of this thesis.

Reasons for this approach can be found in a changing paradigm on the power supply side and in characteristics of data center power demand. To date, the power grid was built to accomodate any power demand from any customer at any time, with the sole exception of emergencies. The result is a planned overprovision in the power grid. In Germany for instance, in 2019, 88GW of conventional and 124GW of renewable energy sites[1] were waiting to supply a demand which in 2019 peaked at around 82.6GW[2] [1]. With a growing political interst to enlarge the share of intermittent renewable energy sources, both the frequency and the amplitude of oscillations in the grid are becoming higher and less predictable. Applying former expansion strategies to the grid would therefore lead to disproportionally reduced utilization rates and at the same time increase investment cost. For these reasons new concepts are needed.

There are many candidates for demand response ranging from people's 'smart' refrigerators via electric vehicle batteries to aluminum production. Recently, data centers have been given a lot of attention as potential participants in de-

[1] `https://www.energy-charts.de/power_inst.htm`, accessed 08/06/2020

[2] These are preliminary data.

mand response schemes: Due to increasingly communication-based production and consumption patterns, they are booming in size, number and power density.

In Germany, for instance, reports indicate that between 2007 and 2017 the overall number of data centers in Germany has increased from roughly 2000 to nearly 3000. Also, within this time frame the number of big data centers has doubled [104]. In the U.S., the development of 'big' data centers is even steeper: Nearly all server shipment growth between 2010 and 2015 was related to hyperscale data centers [183], which will render power draws of over 100 MW per data center more common[3]. The impact of this on the energy and power consumption of the total data center industry is further spurred by an increasing power density inside normal data centers[4][156].

Thus, the digitization of society is resulting in increasing shares of data center electricity consumption at total electricity demand. For Europe, a study commissioned by the EU [36] estimated that electricity used by data centers in 2015 were at 78 TWh, equivalent to 2.5 % of total EU demand. In Germany, in 2017, the overall energy demand from data centers was 13,2 TWh, representing around 2.5% of German final electricity consumption[5]. Globally, power demand by data centers is predicted to represent about 20% of power demand world wide by 2025[6], and in hubs like Frankfurt this percentage is already a fact today[7].

Therefore, due to its sheer size the data center industry is an excellent candidate for demand response. And by their very nature, data centers build on highly automated and often fine-grained computing processes that technically can be tuned to grid requirements in a sophisticated way. At the same time, this enables a high level of automation for implementing demand response schemes at data center sites, should contractual constraints be dealt with. So, not only its size but also the technical characteristics of the data center industry render

[3] http://worldstopdatacenters.com/power/, accessed 08/06/2020

[4] https://www.datacenterknowledge.com/power-and-cooling/new-workloads-cost-pressures-drive-data-center-power-densities, accessed 08/06/2020

[5] calculated based on [105] and [44]

[6] https://www.researchgate.net/publication/320225452_Total_Consumer_Power_Consumption_Forecast, accessed 08/06/2020

[7] https://www.datacenter-insider.de/strom-fuer-die-deutsche-hauptstadt-der-rechenzentren-a-827997/?cmp=nl-86&uuid=00181A4B-B282-4507-B06A3E10CDE5105E, accessed 08/06/2020

it a perfect match for demand response schemes. Therefore, it is not surprising that there is a large body of scientific work examining demand response in data centers. The number of research papers in this field started to grow very slowly after the turn of the century and accelerated around 2010, stabilizing on a high level since around 2015. Most of these acknowledge demand response with data centers as having enormous potential. However, looking into data about demand response with the participation of data centers reveals that there is not much experience with this approach, not even in the U.S., where the concept of demand response was developed and applied long before it became a topic in Europe.

Demand response was first developed in the U.S. as demand side management, i.e. on a mandatory basis and with public intervention rights. In the 1980s a weak power grid was confronted with the formerly unknown load of ubiquitous air conditioning which led to increasing threats of outages. The original idea was that *the utility* could temporarily reduce the load of big power consumers in order to react to temporary problems in the power grid through unexpected increases of power demand [84, 62]. While this concept has been developed and refined to incorporate various scenarios, contractual options and partners, in principle it is a matter of the difference between power and energy: Whenever the instant demand for electrical power deviates from the instant supply, in order to avoid damages to equipment due to electrical imbalances, this gap must be filled, either by supply or by demand flexibility. In the case of demand response, this means that a consumer is required to temporarily reduce or increase their power demand without necessarily changing their overall energy consumption. Whereas energy efficiency projects aim to reduce the energy consumption of processes (i.e. the number of kWh), demand response targets the adaptation of power (i.e. the number of kW) to a temporary problem of size in the power grid. Many industrial processes contain elements that can be temporarily shifted, implying that the *theoretical potential* for demand response is huge. Unfortunately, it can be only partially realized due to economic constraints which turn some technically feasible concepts into economic

impracticality. Therefore, the *economic potential* is greatly reduced when compared to the technical potential. In practice, even economically sound solutions might not be implementable, creating even less *practical potential* for demand response.

However, manipulating power demand to accommodate grid requirements instead of customer requirements may lead to unwanted consequences. In the case of data centers these include increased package round trip times, extended job runtimes or even reduced site accessibility. Also contractual constraints might prevent data centers from touching the operation of their system, or data center management might not be ready for the general concept of demand response. These are a few reasons for the considerable *gap* between the technical potential of data center demand response identified in previous research and its practical implementation, discussed by [217, 28, 37].

1.1. Observations

This dissertation takes a step towards closing this gap by analysing demand response with data centers from a broader and economically motivated point of view. It is based on the following observations:

Despite the large body of research dedicated to demand response with data centers, the real economic potential of data center demand response is only partially represented.

- On the one hand, many research papers deal with very specific scenarios so that the results cannot be generalized and the external validity is low.
- On the other hand, many research papers only address a small subset of flexibility options and associated incentives in a data center so that the flexibility potential of a data center is underestimated.

This partially explains the gap between the theoretical and practical potential of demand response with data centers.

1.2. Hypotheses

In order to avoid the constraints identified through the above observations, the following hyphotheses are made:

- The economic potential for data center demand response can be represented well by a combination of methods that connect an economics-inspired generic framework of demand response with data centers with concrete instantiations.
- The broad view of the framework represents the (technical and) economic potential with a high degree of external validity.
- The specific view of a concrete instantiation represents a high share of the total flexibility of the modeled data center.

1.3. Research Questions

It it the aim of this thesis to support these hypotheses by working on the following research questions:

1. How can demand flexibility in data centers be modeled in order to theoretically encompass power management strategies at all levels of their architecture?
2. How can the high level of abstraction of modeling demand response with data centers be reconciled with technical and economic characteristics of specific data centers?

1.4. Contributions

To answer these research questions, this dissertation will produce the following results, starting with a high level of abstraction that is successively broken down to represent the characteristics of a specific data center:

1. A *modeling framework for demand response* with data centers will be created in the form of multi-strategy, multi-market optimization. It is inspired by micro-economics, and it views the power flexibility of a data center as the *'output'* of a *'production function'* that needs the *'input'* of power management strategies.
2. This framework is validated in a hierarchical approach by first designing a *generic simulation architecture Sim2Win* that models demand response with a variety of different data center types and demand response schemes.
3. In a second step this generic architecture is instantiated into one *specific simulation system Sim2Win-HPC*, which simulates the impact of involving a specific German HPC data center on two German power flexibility markets.

1.5. Structure of the Work

To lay down the contributions of this thesis and how they are being developed, chapter 2, explains the background, introducing data centers, the power grid, and issues involved with demand response schemes. Chapter 3 is dedicated to scientific work related to the presented thesis. It focuses on optimizing and simulating demand response with data centers. As a framework for demand response with data centers is to be designed, this chapter also refers to other research that aims at modeling flexibility of electricity consumption in general. This paves the way for the introduction to the methodology chosen for this thesis and the explanation of the modeling framework in chapter 4. As a first level of evaluation the simulation architecture Sim2Win is illustrated in chapter 5, and

in the second part of this chapter the simulator Sim2Win-HPC is presented in detail. Chapter 6 explains the set-up of the simulation scenario, the planning of the simulation runs, and it documents the results which are discussed at the end of the chapter. Finally, chapter 7 concludes the thesis, summarizing the results and providing an outlook for future work.

2. Background

2.1. Data Centers

There are a great variety of data centers. They not only range in size, business model, and workload, but also in other issues as housing characteristics. These are, for instance, notable in the case of the Barcelona Supercomputing Center which was constructed inside a former chapel (see figure 2.1) .

In order to derive a definition of a data center many authors view data centers from a buildings perspective and focus on the housing aspect of the IT equipment. *Hintemann* [103] or *Barroso and Hölzle* [25] define a data center as 'a building (or buildings) designed to house computing and storage infrastructure in a variety of networked formats.'[25, p.47]. Others provide broad definitions based on the computing infrastructure as *Ghatikar et al.* [89]. As the focus of this thesis is the power flexibility of data centers these definitions fall short of providing a good basis. The requirements for defining a data center in the context of this work are:

- To reflect the overall dependence of a data center's power demand on the workload
- To reflect the impact of operating the physical infrastructure in different modes on the power demand of the data center
- To focus on dynamic, influenceable characteristics of the data center rather than on static characteristics (e.g housing).

Therefore this thesis builds on the definition by *Oro et al.* that encompasses the housing, the infrastructure, and the usage of the DC:

Figure 2.1.: Barcelona Supercomputing Center is constructed inside a former chapel (Photographer: T. Schulze)

> *'A data centre could be defined as a structure, or group of structures, dedicated to the centralized accommodation, interconnection and operation of IT and network telecommunications equipment providing data storage, processing and transport services, together with all the support facilities for power supply and environmental control with the necessary levels of resilience and security required to provide the desired service availability'* [160, p.430].

2.1.1. Power Metrics

From the great variety of data center power metrics only the most well-known and widely used 'Power Usage Effectiveness' (PUE) is shortly introduced. The main reason for its usage is the overall high availability of data relating to this metric. Also it is applied in many data center power models and will here be used in this context. For more information on data center power metrics that