Energy Informatics

RESEARCH Open Access

Spinning gold from straw - evaluating the flexibility of data centres on power markets

Sonja Klingert* and Sebastian Szilvas

*Correspondence:
klingert@uni-mannheim.de
University of Mannheim, Am Schloss, 68131 Mannheim, Germany

Abstract

Data centres have been the focus of research as candidates for demand response or other demand side management programs for quite some time. However, a complete framework optimising data centre demand response is still missing. This is due to the complexity of integrating more than one power flexibility technology and more than one market for power flexibility. In the presented work, this challenge is solved by creating a microeconomics inspired optimisation approach that takes the view of a data centre offering power flexibility as a 'product' to explicit and/or implicit demand response power flexibility markets. This generic framework is then instantiated in a linear optimisation problem that optimises the power flexibility of a German High Performance Computing Centre on a set of different power flexibility markets in Germany. It is consequently shown that, under the described scenario, frequency scaling should be preferred to temporal workload shifting and that the EPEX day ahead market is the most beneficial power flexibility market.

Keywords: Demand response, Data center, Optimization, Power markets

Introduction

In a future with high shares of intermittent energy sources, a power system will need to rely on backup through both the flexibility of power supply and power demand units. Demand Response (DR) is a concept that organizes the interaction between the requirements of the power system and flexible power customers. For quite some time Data Centres (DC) have been discussed as a source of power flexibility, in this work defined as the flexibility to tune the load curve. There are various reasons for that: The load that DCs impose both individually on the distribution grid and collectively on the transmission grid is continuously increasing, rendering DCs a powerful source of flexibility. In Germany, in 2017, the overall energy demand from DCs was 13,2 TWh, representing around 2.5% of German final electric energy consumption[1]. The worldwide data centre power demand is projected to represent about 20% of the overall power demand by 2025[2], and in hubs

[1] calculated based on (Hintemann 2018) and (Bundesverband der Energie- und Wasserwirtschaft 2019)
[2] https://www.researchgate.net/publication/320225452_Total_Consumer_Power_Consumption_Forecast, accessed 08/06/2020

like Frankfurt, this percentage is already a fact today[3]. Also, power density inside the data centre is mushrooming; the largest DCs demand more than 100MW even today[4]. In the case of high and sudden power swings of several megawatts this imposes severe threats to the local distribution grid (Stewart et al. 2019). Technical aspects make DCs perfect candidates for DR concepts: they have highly automated management processes and a fine-grained power load. Many research works see a huge benefit in DR with DCs, both from the DC point of view and for the stability of the grid. However, looking into the empirics of using DC power flexibility in DR or other demand side management programs yields few results (Wierman et al. 2014; Patki et al. 2016). DCs have so far been reticent to engage in DR, except for of some larger DCs sourcing their electric energy on the wholesale market (Patki et al. 2016) and some controlled testing (Ghatikar 2012; Piette et al. 2006).

Data centres demand response

DR in this work is defined as 'voluntary changes by end-consumers of their usual electricity use patterns - in response to market signals... or following the acceptance of consumers' bids ... to sell in organised energy electricity markets their will to change their demand for electricity at a given point in time' (Commission (2013), p.3). This definition by the EU commission allows for the regulation of both increases and decreases of power. DR is both a concept and a practice: as a concept, its predecessor was developed in the 80ies of the last century by a utility that realized the threats of the new load 'air conditioning' to the stability of the grid (Chamberlin and Gellings 1988). Since then it has been implemented through various programs, and the conceptualization of these was introduced belatedly. Markets, where changes of patterns in electric energy use can financially benefit an electricity customer, are called 'power flex markets' in this work. It is helpful to differentiate between explicit and implicit DR (Coalition 2017), also referred to as incentive-based vs. price-based: Explicit means based on a specific demand response contract, whereas implicit DR is the reaction to dynamic prices without contractual binding.

In general, the DR potential can be subdivided into technical versus economic versus practical potential (Jochem et al. 2000; Gils 2014), wherein the economic potential is a subset of the technical potential, and the practical potential is contained in the economic potential. Without accounting for this differentiation, most research works on DR with DCs deal with the technical potential. This, however, is not sufficient for a DC to decide to participate in DR schemes. Several works, for example, suggest migrating workload geographically in response to geographically-differentiated dynamic prices (see "Related work" section), assuming identical nodes and interruptible workloads. Unless a real DC uses virtual machines and has a federated DC with an equally suitable IT infrastructure, unless the target DC has enough space to move the workload into, unless the original DC willingly touches the workload entrusted to it by its customers (which is not the case, e.g. for colocation DCs), and, finally, unless the geographical price difference is high enough to justify the energy and risk overhead, the theoretically identified DR potential crumbles.

[3] https://www.datacenter-insider.de/%20%20strom-fuer-die-deutsche-hauptstadt-der-rechenzentren-a-827997, accessed 08/06/2020
[4] http://worldstopdatacenters.com/power/, accessed 08/06/2020

Because it is reduced to the very specific case of federated DCs under one ownership where migration is in most cases integrated into the business setup as disaster recovery.

Therefore, at least the economic potential needs to be taken into consideration, which however is subject to different regional market aspects and the technical and business characteristics of the considered DC. Even better, the practical potential should be identified which reduces the level of abstraction further by accounting for implementation issues like responsibility settings and general community norms. Among the identified reasons for DCs' low participation in DR schemes, are risk adversity of management personnel (Fernández-Montes et al. 2015; Glanz 2012), organizational barriers inside the DC (Whitney et al. 2014), and obstacles in power regulation such as high minimum thresholds on DR markets (Wierman et al. 2014). One aspect common in all approaches is that they are refined to a specific data centre setting in a specific market. This ergo limits the external validity to highly similar scenarios.

Framework's requirements

To more realistically assess the potential of DR with DCs, a generic framework is missing which does not prematurely exclude specific power management strategies nor power flexibility markets and enables different settings inside the DC. The presented work suggests such a framework that offers these functionalities and enables the instantiation of the framework into a great variety of scenarios.

Creating such a framework demands the following set of requirements be met:

- **R1:** A generic framework should account for both sides of the demand management interaction: Modelling the power flexibility inside the DC *as well as* conditions of the power flex markets[5].
- **R2:** The framework needs to keep the models for data centre flexibility and the market side modelling sufficiently generic to add or remove power management strategies or markets to be tailored to any data centre and environment.
- **R3:** Despite the generalized framework, characteristics need to be extracted and formulated for both sides: Characteristics that are typical for flexibility strategies in DCs and the ones which are typical for power market conditions.
- **R4:** To account for both the economic and the practical potential of DR with DCs, a framework should at least offer starting points to include non-technical issues.
- **R5:** In most cases, a framework on DR with DCs will aim at maximising the DC's benefit; however, optimising other objectives should be enabled. This means that the framework's objective function must be exchangeable; e.g. instead of maximising the DC's benefit, an alternative optimisation function might minimise the draw from the power grid or the system-wide cost.

A theoretical concept that is consistent with these requirements is offered by the microeconomic production theory. It turns one or more 'inputs' into a 'production output' via a technical production function that represents the production process. As far as costs are attributed to the inputs, the production function can be mapped to a cost function. Using several production and cost functions, an optimal aggregated production function and an aggregated cost function can be generated. Meeting with prices on 'markets' (in

[5] Restricting ourselves to 'demand response' implies that we are looking at the market side of the grid and *not* at the physical side of the grid aka smart grid modelling.

this case 'markets for power flexibility', in short, 'power flex markets') the production level is optimised.

Interpreting power flexibility as a side-product of the core business of a DC, i.e. to deliver IT services, allows the modelling of power flexibility as a production process. The 'inputs' are the implemented changes, e.g. shifted workload or a tuned cooling set point; the 'technical production function' is the formal description of the enacted power change and the 'output' is the power flexibility that is offered on the power flex market - dependent on the market prices. This approach makes it easy to add or remove power management strategies customized for a specific DC, these being merely different 'technical production functions', removing respective power flexibility markets.

Our contribution

This work presents a conceptual framework for optimising the benefit of a DC in the context of DR using a microeconomics inspired modelling approach[6]. It allows to combine a set of different power management strategies represented by technical power flexibility production functions and to offer the resulting power flexibility on a set of power markets. Yet unknown power management strategies and power flex markets can thus be easily added. Specific settings in the DC and even non-technical issues can be integrated, e.g. by using constraints. This conceptual framework is evaluated in the form of a linear optimisation problem using a German High Performance Computing Centre (HPCC) offering their power flexibility on German power flex markets as a scenario. Certain Data Centre types are better suitable for certain power management strategies. For a classification of Data Centres and which power management strategies would be applicable, we refer to the work of Klingert (2018). The linear optimisation problem models two power management strategies, temporary workload shifting and frequency scaling, to produce power flexibility. This flexibility is offered on three different power flex markets in Germany: on the EPEX day ahead market as a representative for implicit DR, the secondary reserve market as explicit DR, and a special German solution to power management called 'Atypical Grid Usage' (AGU).

The paper is organized in the following way: "Related work" section, discusses the research basis on which this paper builds, "A framework for modelling demand response with data centres" section, then explains the theoretical concept of power flex functions, cost functions and power flex markets. Subsequently, "The evaluation scenario" section, presents the evaluation scenario and "A linear model instance of Data Centre power flex ibility" section, the linear optimisation instantiation based on the conceptual framework. Finally, "Results" section, presents and analyzes the results of the optimisation and a set of sensitivity runs, and "Discussion and outlook" section, sums up the results and discusses future research.

Related work

This work builds on research in three relevant areas: optimisation of the power flexibility within the data centre, modelling of the interaction with the power markets, and finally the modelling of flexibility in general.

[6]We refer to Wang et al. (2013) for the definition of DR optimisation in the context of DCs. They propose using data centres as a resource for demand response and present a framework to economically optimize data centres' operation for this purpose.

Power flexibility in data centres

Towards the end of the last decade, research on utilizing power flexibility in data centres began separating from energy efficiency-related work. Qureshi et al. (2009) were among the first to give a theoretical foundation of capitalising on geographical and temporal price differences of distributed data centres by routing an interactive load accordingly. Shortly afterwards, in late 2009 and 2010, a team around Ghatikar et al. (2009) at LBNL gave an overview of technologies and discussed general DR opportunities for DCs. They also performed experiments in four DCs, the results of which have been used as a basis in many works.

Technical power management strategies can be grouped according to their level in DC architecture. The categories are as follows: infrastructure-related, hardware-related, software-related, and application-related strategies (Klingert 2018). Most papers on DR with DC are limited to one single power management strategy. An overview can be found in Giacobbe et al. (2015) and Kong and Liu (2015).

An example is dynamic voltage and frequency scaling (DVFS) as a hardware-related energy efficiency strategy which recently began to receive academic attention as a DR strategy. A different example which is intensively researched is workload shifting as a software-related strategy. It is often realized as geographical load balancing, mostly through scheduling algorithms, (e.g. Wang et al. (2012); Liu et al. (2011, 2013); Qureshi et al. (2009) in which the major challenge is to control the increased round-trip times of jobs. Fridgen et al. (2017) use scheduling in geographically-distributed DCs to simulate the benefit to a DC of offering balancing power in a setting where bidding prices are certain vs. forecast, based on real data in an emulated reserve market setting. Another option is temporal workload shifting, that is either pausing jobs and resuming later or scheduling them at a different point in time (Cioara et al. 2016).

Concerning DVFS, most of the works centre around at their effects on *energy* consumption, not specifically considering it for demand flexibility. Exceptions are some works that use this strategy to control peak power, described as power bands in Shoukourian et al. (2015a). Islam et al. engage likewise with DVFS (Islam et al. 2016) suggesting it as a fast responding power management strategy. They show the impact of a market-based approach to coordinate tenants' power demand in a multi-tenant data centre. Wang et al. (2014) see DVFS as one possible workload-impacting strategy which they model as part of a DR framework. This short selection of works, however, is representative for the issue that the external validity of the mentioned approaches is limited, contrary to the suggested framework.

Electricity market modelling

A lot of research focuses on the modelling of the electricity system from a physical or market point of view. Among them are the papers of Zhou and Bialek (2005), who present a model for the European grid system aimed at studying cross-border energy exchanges. In 2017, the European Commission introduced laws for the integration of the European balancing markets ((EU) 2017) to enable energy service providers to offer their services on a European market. Van der Veen et. al. describe in great detail how the European balancing markets work (van der Veen and Hakvoort 2016). For a general overview of the European power sector, we refer to the Agora Energiewende report 'The European Power Sector in 2019' (Energiewende 2020) and for a future outlook to the year 2030 to their

report 'European Energy Transition 2030: The Big Picture' (Energiewende 2019). Currently, there is a European implementation project (entsoe 2019c) for a joint European balancing market in line with the Electricity Balancing guideline with the two sub-projects PICASSO (entsoe 2019b) for automated frequency restoration and MARI (entsoe 2019a) for manually activated reserves. Ventosa et al. (2005) give an overview and classification of modelling techniques of energy markets in general, whilst Haubrich et al. (2001) focus on the transmission limitations of the grid. For general power market modelling, these problems are categorized into exogenous price models (Rajaraman et al. 2002; Fleten et al. 1997; 2002) or models for which the price depends on the decisions of the company (Ventosa et al. 2005). Exogenous price models assume a nearly perfect market and are mostly solved with Linear Programming (LP) or Mixed Integer Linear Programming approaches. The work presented here deals power markets exclusively from the perspective of a DC that considers offering its power flexibility to adequate power markets. Hence, our focus is on the representation of power markets in DR modelling, specifically in DR with DCs.

To deal with flexibility, a DC has three options: It can look into *retail or wholesale electric energy markets* to buy electric energy cheaper. It can turn towards *ancillary or reserve markets* and sell its flexibility there. Finally, it can manage its *power charge*, a case that can be viewed as a long term and static ancillary service. This situation is often modelled as an optimisation problem focusing on one electric energy consuming company that aims at optimising their energy bill.

In most studies where a DC is faced with dynamic prices, these are not derived directly at the wholesale power market. Instead, the dynamic prices are mediated via a utility or power seller and charged in the form of real-time prices (RTP, e.g. GmbH (2018)). The distinction between wholesale prices and dynamic tariffs, also concerning uncertainty, thus is often blurred.

Incentive-based power adoption

In one of the first works dealing with power adaptive DCs, Liu et al. (2011) introduced the concept of geographical load balancing. They discussed using geographical price differences as incentives to adapt scheduling, and thus change the load. In a later work, Liu et al. (2012) schedule according to a dynamic energy price, optimising both cooling and energy storage. Among other works that combined DR of DCs with direct purchases on the power market, are (`?Wang2013`; Rao et al. 2010; Yao et al. 2014). In all these cases, the dynamic prices are modelled as certain or uncertain exogenous price vectors. Mahmud and Ren (2013) wrote one of the few papers that see DCs as price makers, not takers, due to their huge power consumption (assuming 50 MW). They model the price that the DC uses for its calculations, based on a linear regression of historic data, and assume that the DC can reduce its cost by 2%. The current framework model is open with regards to the implementation of implicit DR; the linear model instantiation uses historic EPEX day-ahead prices.

Capitalising power flexibility

The second category of research work deals with DCs selling their flexibility to ancillary services or reserve power markets. Ancillary services are often implemented in the form of peak load pricing, peak pricing or critical/coincident peak pricing (Wierman et al. 2014); the detailed implementations differ significantly from country to country. Liu et al.

(2013), for instance, model coincident peak pricing where peak prices are fixed and the corresponding time slots are unknown. Their model comprises not only peak timing and pricing data, but also algorithms that guarantee a worst-case performance independent of the prediction accuracy. Wang et al. (2012) model a mix of voluntary and mandatory power reductions, wherein they get power reduction signals but must estimate the according price. A model of an ancillary market contract is offered by Ghamkhari and Mohsenian-Rad (2012) who describe the interaction of the grid operator and the DC. They use a frequently-issued request of the grid provider, containing a volume-dependent compensation function to which the DC reacts with an offer of flexibility. This approach is comparable to (Aksanli and Rosing 2014) who model mandatory DR where the DC sends flexibility offers to the utility that has the right to execute adaptation accordingly.

Smart-grid perspective

Some works model the power markets from a physical, smart-grid perspective. Chen et al. (2014) provide an example of this, modelling ancillary service requests based on deviations of frequency from the required value (60 Hz). The DC reacts by scheduling jobs and putting idle servers to sleep. This enables them to capture the DC's power flexibility's impact on the physical grid. Customers can then receive reduced network fees, on condition that they significantly reduce their load in comparison to a year's peak during specific and predefined time windows, which typically change by day and season. The current work to our knowledge is the first to model the secondary reserve market in Germany in the context of DR with DC.

With regards to the third perspective of managing the DC power bill, Xu and Li (2014) present an analysis of the overwhelming share of power charges in 2013 in the US, offering an option to reduce this: partial execution of service jobs like searches, i.e shedding the load, by prematurely terminating the job once the quality of the result is sufficient. Controlling power charges is often implemented as power capping, where a static peak power threshold must not be surpassed (e.g. Shoukourian et al. (2015a) and Berl et al. (2013)). The German Atypical Grid Usage is more complex, but to our knowledge has not been modelled in the context of DR with DCs previously to our work.

Multi-Strategy, multi-Market modelling

The research aimed at optimising more than one technical strategy and more than one market, which is the focus of the provided work, is rare. Liu et al. (2012) optimise the scheduling of workloads in a renewable and cooling aware way. They do not, however, use cooling or the generation of renewable-based electric power as a DR strategy and they do not tackle more than one energy market.

Cioara et al. (2016) explore the options of applying two power management strategies together: workload shifting and manipulating the cooling set-point. Representing the finished EU project Geyser, they additionally suggest a model of an e-marketplace that combines flexibility demand from all potential players into two trading platforms: GEM, a short term e-marketplace, and GAM, a long term e-marketplace. Among others, DCs can access this market place as flexibility-offering units. Contrary to the suggested modelling framework, they limit themselves to assessing the technical potential of DR with specific DCs.

In the case of Tang et al. (2013) who also look into workload shifting and cooling set-point manipulation, the explanatory power of the utilized methodology, a simple regression, is unfortunately limited. In the model of Cupelli et al. (2018) a DC takes part in both an incentive-based and a dynamic-pricing market using three different strategies: the DC can schedule the workload at different times, it can change the cooling set-point and additionally charge or discharge batteries. The two different market options are explored separately, rather than optimised against each other as implemented in the current work. Finally, Wang et al. (2014) provide a framework for all kinds of workload-related strategies that can offer power flexibility into both a dynamic pricing-based market and a peak pricing-based market. They do this by choosing a high level of abstraction that reduces strategies related to workload, to delay, reduce or shed workload or a combination thereof. Like the presented work, they have the goal to 'accommodate a more general set of knobs.' (Wang et al. 2014), p.307. On the market side, they decompose the flexibility demand products into their components 'peak' vs. 'dynamic' pricing, similar to Kirpes and Klingert (2016) who analyse different market components without offering a formal DR model. This work is closely related to the suggested approach, however, contrary to the latter, they do not account for the cost of the degradation of the quality of service.

Modelling flexibility in general

Only lately has the active role of power market participants *offering* power flexibility been seen and modelled. Niedermeier et al. (2016) discuss the advantages of integrated power planning for a load to closely follow renewables' power supply. This integrated planning allows for an optimum of an omniscient planner, but due to the involvement of different entities, it needs to be split into planning (a power plan) and scheduling (any load). They model the flexibility of a load by the cost of three adaptation issues: frequency of changes, size of changes, and notification time.

A different approach is offered by Barth et al. (2018), who try to capture all possible features of flexible loads in a linear optimisation problem. Their work is more specific than Niedermeier et al. in that they model flexibility of loads rather than flexibility dimensions. A definition of flexibility objects in terms of temporal and sizing aspects, so-called flex-objects, is the focus of the work of Šikšnys et al. (2015) who, to trade flexibility on a market, also show how to aggregate and dis-aggregate small flex-objects. Contrary to the presented approach and the definition by Niedermeier et al. (2016), a flex-object comprises the total load to be scheduled and not only the flexible part of it. Therefore, flex-objects can directly be traded on wholesale energy markets but not on reserve markets dealing only with flexibility.

The mentioned works are comparable in that they choose a high level of abstraction as in the presented framework. However, contrary to the presented approach, even though these are highly interesting analyses of the technical DR potential, they fully ignore the economics of the decision basis.

A framework for modelling demand response with data centres

As mentioned in the Introduction, a generic economics-inspired optimisation approach can be used to meet the requirements R1 to R5. A framework built on this idea will be presented in the following sections.

This generic framework takes the view of a DC that aims at *'selling'* its power flexibility or using it to optimise energy purchase, asking in turn for compensation that more than outweighs the invested effort. This perspective can be understood considering Amazon's situation previous to inventing the cloud. Until Amazon explored selling unused capacities, they spent eleven months of the year suffering from the huge block of fixed cost due to servers sitting idle. Only when they discovered their huge flexibility potential, did they invent the cloud. Using production functions to represent this point of view, adding or removing power management strategies is simply a question of adding or removing 'products' in the 'aggregate power flexibility product function' (in short 'power flex function'). As will be shown, this modelling approach allows the inclusion of non-monetary issues at the constraint level. Once the framework is instantiated to a set of specific strategies and markets, it incurs an optimisation problem for which the most straightforward methodology is linear or non-linear optimisation. Simulation can be used for approximating exact solutions.

Production of power flexibility

Micro-economic enterprise theory views an enterprise as an organizational unit that creates a product or a set of products (called outputs) using one or more resources (called inputs) and a production technology which is represented by a production function. Producing evokes cost; the products are priced according to a specific market structure. In the case of perfect competition, an optimum is reached when marginal costs equal marginal revenue (Mas-Colell et al. 1995).

Translating this idea into a DC offering power flexibility to a set of power flex markets, the product becomes the aggregated power flexibility which is produced applying one or more 'power flexibility production functions' (aka 'power flex functions'). Each of these uses one or more 'inputs' which are characterized by the necessary tuning inside the DC. The more inputs are used, the more power flexibility is created until maximum capacity is reached. In the framework, power production using one specific or an aggregation of several power production technologies is therefore expressed as a (linear or non-linear) function that can be both positive and negative. In the case of power reduction, the flexibility offered to the market is positive. For example, if due to excessive supply of solar power, the DC is asked to reduce its power, this power flex function is negative. The power flex function is continuously increasing (or decreasing if it is negative), using inputs that can be continuous or discrete variables. Mathematically speaking, we have a set of power production strategies $S = \{S_1, ..., S_n\}$ and a set of production inputs $I = \{I_1, ...I_m\}$. These are combined to produce power flexibility as

$$PF = PF(y_{s,i}) \text{ with } s = 1...n \text{ and } i = 1...m \tag{1}$$

where $y_{s,i}$ is the power flexibility output of technology s at the input level i. It continuously increasing:

$$PF(y_{s,a}) \geq PF(y_{s,b}) \text{ if } i_a \geq i_b \; \forall PF \geq 0 \tag{2}$$

$$PF(y_{s,a}) \leq PF(y_{s,b}) \text{ if } i_a \leq i_b \; \forall PF \leq 0 \tag{3}$$

It may have increasing, decreasing, or constant returns to scale. The shape of the power flex function depends on the underlying power management strategies and how they are interrelated.

As a concrete example for this concept, take the two power flex production functions of 'workload shifting' (index $_{shift}$) and 'frequency scaling' (index $_{FS}$), which are implemented in the linear optimisation model instance (see "A linear model instance of Data Centre power flexibility" section):

The background is a widely cited server power model based on CPU frequency (Elnozahy et al. 2002), $P_{serv} = P_{idle} + A * f^3$, where P_{serv} is idle server power and $A * f^3$ is the dynamic part of the server, P_{dyn}, impacted by CPU frequency f and a factor A that includes application-specific and server-specific characteristics (for more information on the power model, please refer to (Elnozahy et al. 2002) and (Shoukourian et al. 2015b), amongst others). The workload in this example is assumed to be defined by j uniform jobs that can be interrupted at any time and use one or more identical computing nodes to create the DC's services. DC power P is assumed to be determined by the server power P_{idle}, by the number of active and idle nodes N, the dynamic power P_{dyn}, and the number of active nodes n. The dynamic average power per job P_j is then determined as $P_j = \frac{n*A*f_0^3}{j}$.

The workload shifting power flex function builds on the changes in the number of active jobs at any point in time, and can be described as $PF_{shift}(\Delta j) = \Delta j * P_j$, with the capacity constraints $0 \leq |\Delta j| \leq \hat{J} - j$ and $j + \Delta j \geq 0$, where $\hat{J}$ is the maximum shiftable workload in terms of jobs. The change in jobs Δj is enabled only between the number of currently running jobs j and the maximum number of jobs $\hat{J}$. This can be determined through technical or business model constraints, or even due to general risk adversity. The independent variable, i.e. the input into this power flex function, is the shifted workload Δj, which is reflected on the x-axis. As we assume uniform jobs, the function in this example is linear, which can be seen in Fig. 1. The resulting Power Flex can be seen on the y-axis. It is based on the average values of a German HPCC. This DC on average runs 160 jobs; assuming that 12,5% cannot be touched, the capacity constraint is 140 jobs which are moved if the reward is high enough.

Likewise, a power flex function can be created by scaling the frequency. This means that the power flex production function PF_{FS} uses the distance to the original frequency f_0 as

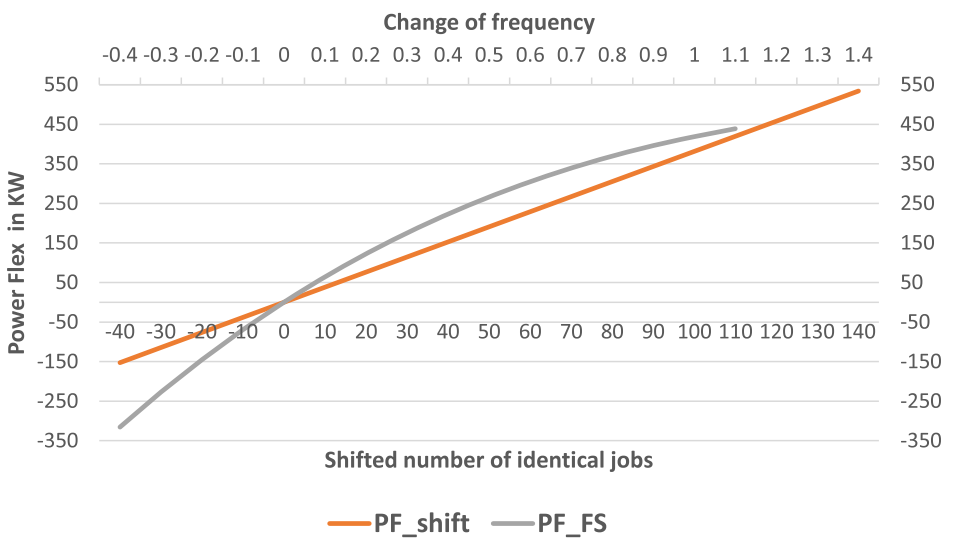

Fig. 1 Power-flex functions of shifting and frequency scaling

input, which is the independent variable for the power flex function. As the frequency is an input for the original server power function to the power of 3, the resulting function would not be linear: $PF_{FS}(\Delta f) = n * A * \left(f_0^3 - f_1^3\right)$, with$f_1 - f_0 = \Delta f$, where f_0 is a parameter, whereas Δf is the independent variable.

In reality, the frequency cannot be manipulated continuously, but there are several technically-possible frequency levels (also called P-states). To linearize the frequency scaling power flex function the power consumption of the workload is normalized to the default frequency f_0. The power consumption in all other cases concerning manipulating frequency can be calculated using the above formula (of course calibrated with the data at hand) and then scaled accordingly. Thus the allowed solution space for PF_{FS} is pre-computed as:

$$PF_{FS}(\Delta f) = \begin{cases} PF_{FS}(\Delta f = f_0 - f_{min}) = n*[P_0 - P_0 * (1 - a_{min})] \\ \vdots \\ PF_{FS}(\Delta f = 0) = 0 \\ \vdots \\ PF_{FS}(\Delta f = f_0 - f_{max}) = n*[P_0 - P_0 * (1 + a_{max})] \end{cases} \tag{4}$$

where a_{min} is the scaling factor for the minimal and a_{max} for the maximal frequency value. PF_{FS} can become negative too when the power flex market requests for a power increase. In Fig. 1, the values of PF_{FS}, based on the data of a German HPCC, are interpolated. In this case, the default frequency is at 2.3 GHz; it can be scaled down to 1.2 GHz and up to 2.7 GHz.

To aggregate the selected power flex functions from the different strategies into an aggregated power flex function that depends on all inputs, interdependencies of the strategies need to be taken into account. For the case of the specified shifting and frequency power flex functions, it is obvious that at a certain point in time, the results of frequency scaling can be attributed to the power flexibility only to the degree that the workload has not been shifted away. This means that shifting workload at the same time reduces the potential of power flexibility that can be reached through frequency scaling. This might seem rather obvious but combining more strategies will render these interdependencies very complex. So, sticking to the simple example, the aggregated form of the power flex function is $PF(PF_{shift}, PF_{DVFS}) = PF_{shift} + \frac{j_0 - \Delta j}{j_0} * PF_{DVFS}$. This can be either positive (i.e. reducing power consumption) or negative (i.e. increasing power consumption), depending on the flexibility the power markets demand.

Quality reduction of power flexibility

An implicit assumption of DR is that power demand is mainly temporarily changed, i.e. the core business of the data centre is not touched, and therefore the size of the workload measured in the number of tasks, services, or other non-energy related metrics remains constant system-wide. What might change, however, is the quality of service related to the service creation, depending on the power management strategy applied. The quality of service impact of a power management strategy can be generally expressed by a set of quality reduction parameters $QR = \{QR_1, ..., QR_l\}$, each of them being subject to change, depending on the strategy applied and the degree to which inputs are employed:

$$QR_r = QR_r(y_{s,i}) \text{ with } s = 1...n;\ i = 1...m \text{ and } r = 1...R \tag{5}$$

with the characteristics:

$$QR(y_{s,a}) \geq QR(y_{s,b}) \text{ if } i_a \geq i_b \; \forall \; QR \geq 0 \tag{6}$$

$$QR(y_{s,a}) \leq QR(y_{s,b}) \text{ if } i_a \leq i_b \; \forall \; QR \leq 0 \tag{7}$$

Again, this quality reduction function may be positive or negative and is continually increasing in its determinants. A specific case of quality reduction is the delay, which is also the most frequently used characteristic. It is necessary to be spelt out because the delay may have double functions: on the one hand, it is a quality reduction characteristic, depending on the technical power adaption strategy used. On the other hand, it is intertwined with the temporal requirements by power flexibility markets to hold up the power adaptation.

For the examples introduced above delay is the only dimension of quality reduction considered. For both the shifting as well as for the frequency scaling power flex function, the corresponding delay can be calculated quite easily. In the case of shifting, the delay is just the time for which that workload is shifted away: D_{shift}, implying that the deadline is identical with the original expected finishing time of the execution. This is determined by the shifting duration that the DC management chooses, which is impacted by power market requirements. In the case of the frequency scaling power flex function, the easiest way to calculate delay is by assuming that the workload is compute-bound only. In this case, the technical delay is determined exclusively by the relative increase or decrease of frequency. $D_{FS}(\Delta f) = -\left(1 - \frac{f_0}{f_0 - \Delta f}\right) * ET$, where ET is the original execution time. If the workload is both compute and memory-bound, the impact of frequency scaling on delay is reduced by a factor β. This represents the relative shares of compute vs. memory-bound execution (for more information, see the work of Etinski et al. (2012) and Auweter et al. (2014)) . These examples show that even though the quality reduction depends on the strategy employed, it needs not to have a direct relation to the shape of the power flex function. Quality reduction itself has no effect on the benefit of marketing flexibility for a DC - the real effect is mediated via cost which will be explained in the next section.

Cost of power flexibility

The cost of power flexibility is modelled alongside the aggregation of three basic elements:

- quality reduction C_{PF}, which ultimately depends on the technical characteristics of the power flex production through the power management strategies,
- fixed cost of each strategy $c_{fix}(s)$, where applicable (e.g. cooling manipulation might require human intervention),
- changed power cost C_{PC} as by increasing or reducing power, the power charge of a bill might be changed[7].

As in the cases of production and quality reduction functions, the high level of abstraction allows for either just a parameter or a separate function that might depend on the level of technical power flexibility. Therefore these costs are represented by $C_{PF} = C_{PF}(y_{s,i})$ with $s = 1...n$; $i = 1...m$, with the same characteristics as the equations for the power flex and the quality reduction functions 1 and 5. The function depends on the technical power flexibility, even though it is mediated by quality changes. As mentioned,

[7] When time aspects are added to the picture, additionally the change in energy cost must be accounted for.

it is not quality reduction per se that determines the impact of a power management strategy on the economic benefit of a data centre, but rather the associated costs. This means that implementing a power management strategy, therefore, might have different kinds of consequences, that are not reflected in costs. In some cases, (e.g. manipulating the cooling set-point) there will be no quality reduction of a strategy as long as it is applied within allowed and customized capacity levels. In other cases, there might be a Quality of Service (QoS) impact, but unless there is an SLA contract, i.e. a service level agreement that determines specific service features that turns this impact into cost, it is not relevant for a power flex cost function (e.g. delay within SLA-boundaries). In yet other cases, there might be no contract-based cost of a power management strategy. Nonetheless, DC management may have other reservations about applying a specific power management strategy, due to general norms in the community, personal interests, or the fear of losing customers (e.g. 'we must have high server utilization, we would never reschedule the workload'). These kinds of non-financial costs can be reflected by constraints in the framework. Quality might not only be reduced but also increased, e.g. if the workload is advanced due to the high availability of renewable power sources. This would normally carry neither costs nor rewards. In the case of a GreenSLA concept (Basmadjian et al. 2016), however, the support of DR by customers might be rewarded which would be reflected by a change of arithmetic signs.

All these cased are captured in the model due to the generality of the approach.

As to fixed cost, they are reflected by a parameter $C_{fix} = \sum c_{fix,s} * b_s$ that needs to be included into the overall cost function, where $c_{fix,s}$ are each strategy's fixed cost and b_s is the corresponding Boolean. Fixed cost act as a threshold for applying a specific power management strategy, as they must be compensated before the specific strategy is beneficial. Automation, e.g. connecting to a DCIM (data centre infrastructure management) tool, can help to reduce the fixed cost.

By the nature of power flexibility strategies, they change the power used by a DC. Depending on the energy tariff, in most cases, this does not lead to a reduction of the power charge as long as the original peak power is not changed. If, on the other hand, the power flex function is negative, i.e. by increasing the power demand, the power charge might change if the original peak power is surpassed. Therefore the additional power charge of the aggregated power flex function of the data centre is described by the constraint $C_{PC} \geq pc*(P_0+PF-\hat{P})$ and $C_{PC} \geq 0$. C_{PC} denotes the change of the power charge, pc the power charge per kW, P_0 the current and $\hat{P}$ the maximum power of the considered billing period. The additional constraint $C_{PC} \geq 0$ ensures that this applies only if the peak power $\hat{P}$ is surpassed. In the unlikely case that the tariff does not have a power charge based on a threshold, the constraint can be simply omitted. In the same way, energy costs can be included in time-dependent instantiations of the model, if the tariff is in kWh.

The total cost C is then just an aggregation of the cost elements:

$$C = C_{PF} + C_{fix} + C_{PC}. \tag{8}$$

In the scenario used for the linear optimisation problem, unfortunately, the DC presented does not have a publishable Service Level Agreement cost model. Therefore, the SLA model for non-interactive batch jobs of Garg et al. is used, assuming that the deadline is at the end of the execution time (Garg et al. 2014). SLA cost is then derived proportionally to the delay: $C_{PF} = D * Pe$, where D is the delay introduced above and Pe the penalty

rate per time step, as the workload is measured through job energy. Three of the leading online services for computations, Microsoft, Amazon, and ScaleMatrix, all define their SLAs dependent on the usage price (Services 2018; ScaleMatrix 2018; Microsoft 2018). This is the reason for SLA costs to be assumed as $Pe = uniform(l_l, l_u) * uP$. The usage price SuP relates to the workload size; the time dimension is included linearly as the SLA cost is computed for each time slot separately.

Modelling power flexibility markets

In most cases, the Data Centre is a price taker on a regulated or a competitive market for power flexibility. In some specific cases of explicit DR, the DC bids for prices; however, as this process precedes the internal optimisation (sometimes by more than a day) it is not integrated into the framework. The power flex markets are therefore represented by their respective prices (price vectors in the time-dependent linear optimisation model) and typical constraints.

These constraints can be rather limiting for the options of DCs to trade their power flexibility on the power flexibility markets. This applies foremost to explicit DR schemes. Implicit DR schemes, i.e. dynamic prices via dynamic tariffs, or a wholesale market like the EPEX spot market in Europe, regularly have far less restrictive entrance barriers. This also applies to time-of-use tariffs, which, not being dynamic, are not DR but help to reduce the pressure on the grid in times which are generally considered difficult. The reason is the nature of explicit DR; it is aimed at physically securing overall power system stability, contrary to implicit DR which is market-based.

For implicit DR therefore, the reward for flexibility demand is modelled via the difference between the baseline power price and the dynamic power price Δp. Explicit Demand Response markets, e.g. capacity markets or reserve markets, offer a reward for either any amount of power (in kW) Re or for each package p^Z of flexibility. On the French capacity market, for instance, only specific amounts of power can be offered through certificates, which are then reimbursed in terms of the number of certificates traded. This can be introduced by defining a variable z as the number of certificates of size Z offered $z = \frac{PF^e}{Z}$ for the power flexibility PF^e sold to market e[8]. As a constraint, this number of certificates sold must then be natural $z \in \mathbf{N}$. The turnover on all k markets $e = 1...k$ comprising all options of explicit and implicit DR is then:

$$T = \sum T^e = \sum (p_{Z^e} * z^e + Re^e * PF^e + \Delta p^e * PF^e) \text{ with } e = 1...k \tag{9}$$

with the corresponding constraints:

$$\begin{aligned} PF^e &\geq \check{M}^e * b^e \\ C^e &= \sum c^e * b^e \\ bigM * (1 - b_s^e) + D_s &\geq D^e \; \forall \, s = 1...n, e = 1...k \\ \hat{M}^e * b_s^e \geq PF^e &\geq \check{M}^e * b_s^e \end{aligned} \tag{10}$$

The first constraint is typical for both explicit and implicit DR and ensures that the DC's offer equals or surpasses the minimum power $\check{M}^e$ that can be traded on the respective market e which is enabled via the corresponding boolean b^e. In the second constraint, fixed market entrance costs C^e are formulated in a way comparable to the fixed costs of a

[8] To maintain a better overview which variables or parameters are market-related vs. data centre-related, the index for the market is kept as a superscript instead of a lower script.

power management strategy in a DC. The third constraint deals with the required adaptation times. In explicit power flex markets, as e.g. ancillary services markets, the market operator offers a reward for power adaptation PF^e which is defined through notification time, frequency of change, adaptation size, (Niedermeier et al. 2016) and the required adaptation duration D^e for market e. This means that if some part of the flexibility created by strategy s is offered in market e, the equivalent delay D_s (activated through the corresponding boolean b_s^e and the mathematical $bigM$) must exceed the delay time required by the market D^e. The behaviour of this boolean is defined in the last constraint; as soon as a strategy s delivers power flexibility to the market e, the total power flexibility PF^e must be within the minimum requirement $\check{M}^e$ and the maximum requirement $\hat{M}^e$ of the considered market e.

As an example for an impicit and an explicit power flex market, take the European stock exchange EPEX and the German secondary reserve market, which are modelled in the linear optimisation instance of this framework and will be introduced in more detail in the evaluation part of this work (A linear model instance of Data Centre power flexibility). Regarding EPEX, the reward is expressed as price difference Δp between the EPEX prices and the baseline tariff with a minimum offer constraint M^{epex}. The Secondary Control Reserve market has minimum offer requirements M^{res} and high market entrance costs c^{res}. The benefit is modelled alongside the difference between the offered load (which carries a specific price per kW) and the number of time periods (z) that this load is activated - mandatory, in case the bid of the DC is accepted in a specific week.

Optimisation objectives

In this framework, the overall optimisation objective is a matter of the use case of the implemented version. Reconnecting to the original microeconomics inspired approach to modelling, it is, of course, the benefit of the company, aka DC, from engaging in DR that should be maximised. However, in some cases, especially in the long run, it might be possible that the DC aims at being a 'good citizen'. A survey of some 20 publicly-owned DCs unearthed that this was their guiding principle for the interaction with their grid and electric energy providers, due to a lack of direct DR products (Patki et al. 2016). To reflect this in the modelling framework, the economic benefit of the DC might be turned into a constraint, requiring that benefit needs to be at least non-negative. The optimisation goal might then, for example, be to minimise the difference between an exogenous power parameter (vector) determined by the grid operator and the overall power (as $P_0 + PF$) as suggested in the general DR framework of (Barth et al. 2018). Usually, however, for a data centre to get engaged in new markets and offer their power flexibility, they would want to maximise their economic benefit:

$$max \left(\sum T^e - \sum C^e - C \right) \tag{11}$$

This is also the objective function used in the linear optimisation instance as presented in the subsequent chapter (The evaluation scenario).

Accounting for non-technical issues

The framework offers three starting points for integrating non-technical issues:

- A solution that is chosen often in the context of micro-economic modelling is to use alternative objective functions, for instance maximising the utility of stakeholder or the system as a whole. Strictly speaking, this is not covered in the framework but requires slight adaptations of the model and a set of new constraints.
- Another option is to integrate an alternative cost function that reflects the apprehension of the involved personnel. There is evidence that also in the context of data centre management decisions on energy-saving measures are not only based on economics but also influenced by apprehension (e.g. Fernández-Montes et al. (2015)). An alternative cost function could be used to compare the optimisation results of both cost functions. This seems very straightforward at first, and this approach was used in the first version of the model (Klingert and Becker 2017). However, there is no data to realistically model this.
- A third solution is to monitor the economic benefit of the optimisation and compare it to reality where ever possible. The difference between the realized benefit and the calculated benefit then needs to be analysed to find out the origin.

The reasoning behind the third option is the following: for a DC manager, some strategies might 'feel' riskier than others due to perhaps complexity (for example, if various different technologies are involved in a cooling setup) or customer issues (in the case of workload-related strategies). For the decision-maker, such a 'risky' strategy must then be a lot more beneficial than a 'safe' power management strategy to be implemented. This would imply that the *required* benefit of the 'risky' strategy must be greater than the one of the 'safe' strategy. Am implementation of this approach is the calculation of the ROI (return on investment) per strategy:

$$ROI_s = \frac{\sum T_s^e - C_s - \sum \frac{T_s^e}{\sum T_s^e} * c^e * b^e}{C_s + \sum \frac{T_s^e}{\sum T_s^e} * c^e * b^e} \tag{12}$$

The benefit related to the turnaround of a strategy in all markets $\sum T_s^e$ must be related to the strategy's cost C_s and the share of this strategy at the fixed costs in the respective markets $\sum \frac{T_s^e}{\sum T_s^e} * c^e * b^e$. If this ROI is considerably higher than the standard ROI of other business engagement in this particular market, the reasons should be analysed.

With financial data missing for the evaluation scenario, however, none of these options could be realized in the linear optimisation instance.

The evaluation scenario

The general framework is evaluated using data from a German HPCC and a set of German power flex markets. The European setting is chosen to support a new geographical perspective on DR with DC as the research in this area is still dominated by U.S. scenarios.

Data Centre traces

The data traces for the evaluation of the framework are provided by a large scale HPCC system in a German DC with more than 9000 compute nodes, each of which features 2x8 core Intel Sandy Bridge processors. They each have a thermal design power of 130 W. Default CPU frequency is set to 2.3 GHz, the maximum operating frequency is 2.7 GHz. The total energy consumption in 2014 was around 2 GWh; the theoretical peak power is near 4 MW. The DC computes scientific workload consisting exclusively of batch jobs.

The data traces provided contain job data with starting and finishing time, energy and average power consumption of each job in 2014. After data cleansing, almost 400,000 jobs are left. Since the DC is an operating environment, the origin of the data cannot be disclosed.

German electrical power markets

Baseline tariff

In Germany, it is typical for the industry to have either a fixed price or a simple, two-period based time-of-use tariff. As the real tariff of the DC at hand is not known, average industry prices from (Bundesnetzagentur 2014) are taken as fixed price tariff.

EPEX spot market

The European Power Exchange (EPEX) is the main wholesale market for electric power for German customers. It consists of several sub-markets: Day-ahead-auction, Intraday Continuous, Intraday-auction markets and the capacity market. This research work considers the Day-ahead-auction market, where energy can be traded in standardized blocks, consisting of several hours or hour blocks. The trading prices and corresponding volumes are published on the EPEX website[9]. To trade on the EPEX, market participants have to pay an entrance fee and a trading fee, and they have to be able to at least trade 100 kW.

Reserve markets

In Europe, reserve markets are distinguished into primary, secondary, and tertiary reserve markets who offer ancillary services by way of passing work from one to another. The secondary reserve takes from the primary reserve resource, and later on, the tertiary takes over from the secondary reserve. The secondary reserve market is the most attractive one for DCs; the rewards are comparably high, but the requirements to adapt power within 5 min can be fulfilled by most power management strategies. Unfortunately, the pre-qualification effort and cost are quite high, and the minimum size of adaptation of 5MW (in 2014) cannot be reached by most German DCs so that for the current scenario it must be assumed that an aggregator mediates between the secondary reserve market and the HPC DC.

There are four products in the German secondary reserve market. They are defined along with the dimensions ‘high and low tariff’ as well as ‘positive and negative reserve power’ which is equivalent to a positive vs. a negative power flexibility function.

The market for Secondary Control Reserve is an auction with two merit-order lists into which the DC bids for each week separately. In the first merit-order list, all auction participants are sorted by their load price in ascending order. The necessary amount of Secondary Control Reserve for the upcoming week is calculated based on a mathematical formula. Participants are accepted into the pool until the predicted load limit is reached. Next, during the week, participants are retrieved based on the second merit-order list, which sorts the participants’ energy price bids. Starting with the lowest energy price, the participants are requested to deliver their service until the demand for SCR is satisfied and are paid according to their bid. Therefore, for each of the products and in each week a pre-qualified market participant (e.g. a DC) offers a combination of a price for power (kW) and electric energy (kWh). Since a Data Centre can’t know exactly

[9] http://www.epexspot.com/en/market-data/dayaheadauction/chart/auction-chart

in which week they will get into the participant pool and how much their service will be used, we need to determine a realistic strategy. From historical data, one can derive that with 80% of the price offer, it is possible to get into the participant pool in 90% of the weeks. Among the providers that got into the pool, one can then calculate the percentage of the previous week's energy price, which led to the highest revenues. To avoid additional computational weight by adding a further optimisation problem to the current problem, this market participation is modelled exactly according to the strategy described above outside the DR optimisation and thus turned into a parameter vector for each week.

Atypical Grid Usage

Atypical Grid Usage is a German peculiarity based on a power charge policy that aims at flattening the aggregate power demand profile in the transmission grid. It uses information about the historically known peaks transmission grid load. Based on this, so-called 'peak load windows' are determined wherein the power consumers are rewarded if they reduce their load significantly compared to their 'regular' load. This means that $PF_{atyp} \geq 0$ should be maximised during these peak load windows, reducing the peak power at these times as much as possible. The network fee itself is not changed, however, it is recalculated in a way that it is only paid for the maximum power in the peak load time windows. Constraints are a minimum load shifting potential of currently 100 MW, a bagatelle threshold for the reward and a so-called 'materiality' threshold of the overall network fee (der Justiz und Verbraucherschutz 2018).

A linear model instance of Data Centre power flexibility

The power flexibility framework is evaluated using a time-dependent linear model on the batch workload data as described above. The implemented power adaptation strategies are workload shifting and frequency scaling.

The model has the following assumptions:

1. The workload is composed of several uniform tasks without user interaction, which can be interrupted and restarted at any time, which implies a virtualization layer.
2. The workload is represented by the amount of dynamic server power necessary to compute the tasks.
3. Both power management strategies, frequency scaling and workload shifting can be implemented without delay and without a significant performance reduction (Virtual Iron Software 2007).
4. The workload shifting is modelled as postponing workload by a maximum number of time steps (the so-called flexibility range)
5. Since load is always only measured at certain time steps, the power consumption in between has to be deemed as constant. The power values (W) are converted into energy values (Wh).
6. Even though electric load might be shed due to frequency scaling, which changes computational efficiency, workload itself is not being shed. Accordingly, the processed workload in Wh is normalized to a standard frequency remains constant.
7. The quality reduction is viewed in terms of delay, to which SLA costs are attributed.

8 The DC power consumption is abstracted from other power consumers than server power consumption as only the workload and CPU-related power management strategies are implemented.

As Fig. 2 shows, the linear optimisation model is composed of four interdependent sub-models: the technical power flex functions component, the DC cost component and the delay that links the other two. And finally the power flex markets component with one sub-component each for each power flex market. Adaptation processes lead to an adjustment in the DC in the form of power flexibility originating from the two power management strategies, which impact the quality of service in the form of a considered run-time model and thus connects to the cost model for SLA. On the market side, both direct and indirect demand response is modelled: Indirect DR is represented by the EPEX day ahead market, whereas direct DR is represented by a model for the German Secondary Control Reserve market as well as for Atypical Grid Usage introduced in "The evaluation scenario" section. The various components are interdependent: Reward structures and constraints on the power markets have an impact on the power flexibility offered, which in turn is dependent on the internal quality of service characteristics and cost structures.

The following sections explain the modelling approach to the linear optimisation model, first with regards to adaption processes inside the DC and then with regards to the considered power markets.

Data Centre models

Power flexibility through workload shifting

According to the linear power model for workload shifting as introduced in the section 'A framework for modelling demand response with data centres', the power flexibility of workload shifting is simply dependent on the number of jobs j and their average job power P_j: $Pf_{shift} = \Delta j * P_j$.

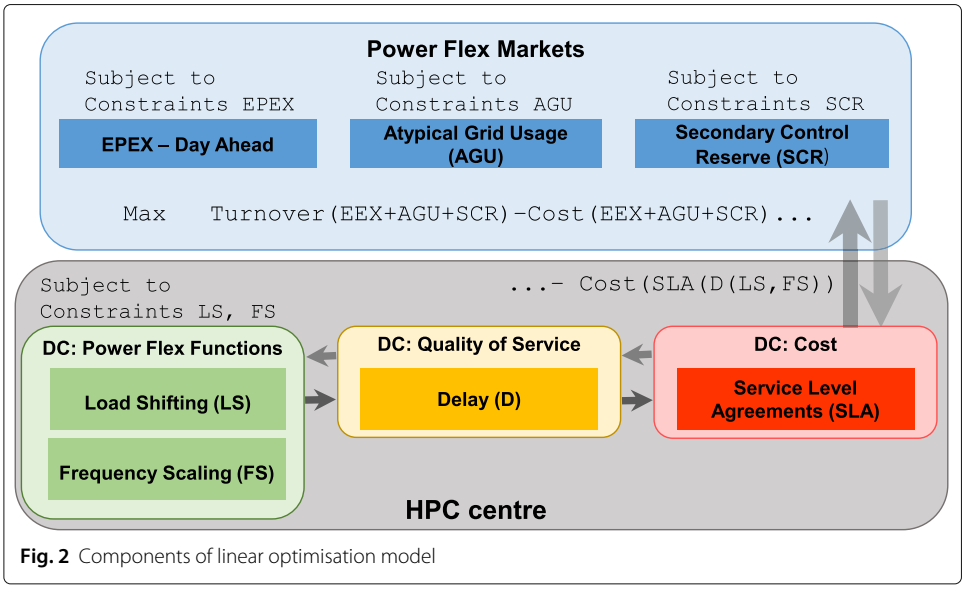

Fig. 2 Components of linear optimisation model

The current model implementation is time-dependent, which means that the power flexibility function carries a time index and shifting is carried out from one period to another. The set of periods of the considered time range is TR where one period is the interval between two measurement points. The workload is shifted from originating periods $o \in TR$ to one or more feasible resulting periods $r \in TR$. All combinations would be a tuple set of $TR \times TR$. However, only the tuples inside the flexibility range FR, defined as $FR = \{t_{min}, ...0, ..., t_{max}\}$ are valid. $t_{min} \leq 0$ is the maximum possible number of periods to move jobs back in time and t_{max} is the maximum number of periods to move jobs forth in time. Thus the set of valid tuples U is defined as $U = \{(o, r) \in TR \times TR : r - o \in FR\}$.

As an example take the set of periods $TR = \{1, ..., 4\}$ with a flexibility range of $FR = \{-1, ..., 2\}$. The viable combinations U are shown in Table 1.

As workload o_o (the amount of energy o used for jobs in period o) is represented by the energy used to compute it, shifting is done by moving the energy amount of every originating period $o \in TR$ to every resulting period $r \in TR$, where $r - o \in FR$. In the example in Table 1 the energy of originating period $o = 2$ could be shifted to the resulting periods $r \in 1, ..., 4$. For the originating periods 1, 3 and 4, the flexibility range FR cannot be fully utilized, as the resulting period r must be within TR.

As the jobs are expressed via their consumed energy, the (non-negative) decision variable Δj in the power flex function is simply the share of energy moved from the period o to a viable period r. The sum of these shares (summing up along the rows in Table 2) must be equal to 1: $\sum_r d_{(o,r)} = 1, \forall\, o \in TR$, where $d_{(o,r)}$ is the share of demand shifted from the originating period to the resulting period. The absolute magnitude in the power flex function is determined by multiplying this share $d_{(o,r)}$ with the original energy consumption in the originating period and summing up the resulting values (along with the columns in Table 3), which is expressed in the constraint: $e_r = \sum_{o \in U} d_{(o,r)} * o_o \forall\, r \in TR$, as can be seen in Table 2, where o_o is the original energy load in period o and e_r[10] is the resulting energy load in period r. The difference between the original load o_o and the remaining energy load e_r in period r is then the power flexibility of shifting PF_{shift}.

Finally, the following constraint ensures that the optimisation model never exceeds the maximum load of the data centre, which could be shifted to a period $r \in TR$, denoted by $\hat{J}_r$: $e_r \leq \hat{J}_r \; \forall\, r \in TR$. Additionally, $e_r \leq \hat{E} \; \forall\, r \in TR$ is the overall capacity constraint, where e_r again is the energy load in the resulting period r and $\hat{E}$ the maximum capacity of the DC. In the example (2), the price difference Δp between the current cost and investing into the day-ahead market comes into play: By moving the load from period 2 into periods 3 or 4 an additional revenue can be made. Let 40 kWh be the overall capacity constraint and 50% the maximum shiftable amount in any period. The original amount of job energy in period 2 (second row) is 40 kWh. Even though the price difference Δp between 2 and

Table 1 Tuple set U of originating periods o to resulting periods r

r / *o*	**1**	**2**	**3**	**4**
1	1,1	1,2	1,3	
2	2,1	2,2	2,3	2,4
3		3,2	3,3	3,4
4			4,3	4,4

[10] Here e is an energy value; not to be confused with the market index ^{e} in the framework model.

Table 2 Results of shifting the shares $d_{(o,r)}$

r / *o*	**1**	**2**	**3**	**4**	o_o	Pf_{shift}
1	50%	0%	50%		10	+5
2	0%	50%	12.5%	37.5%	40	+20
3		0%	100%	0%	30	-10
4			0%	100%	20	-15
e_r	**5**	**20**	**40**	**35**	100	
Δp	1€	2€	5€	4€	*kWh*	

e^r is the resulting energy load!

3 is most profitable, due to the capacity constraint only 12.5% can be shifted from period 2 to period 3, thus resulting in a job energy consumption of 20kWh in period 2 and 40 in period 3. All in all, shifting the 50% out of period 2 leads to a power flex value PF_{shift} of +20 for period 2, which means that 20 kWh can be offered on the day ahead market.

Power flexibility through frequency scaling

The second power management strategy implemented in the linear instantiation of the power flex framework is frequency scaling. As explained in "Production of power flexibility" section, in this model the power flexibility function by frequency scaling is determined by a linear power factor depending on the selected change of the normalized frequency. This means, that the finally chosen frequency $f \in F$ in the following constraint is determined through a boolean which makes sure that only one frequency is chosen: $\sum_{f \in F} b_{f,r} = 1 \forall r \in TR$.

Thus, taking the frequency scaling strategy into account, the energy in a resulting period r can be described as $e_{f,r} \leq \sum_o d_{f,(o,r)} * o_o \forall\, r \in TR, f \in F$, where $f \in F$ is the frequency chosen through the boolean variable. This is realized by $e_{f,r} \leq BigM * b_{f,r} \,\forall\, r \in TR, f \in F$, and $e_{f,r} \geq \sum_o d_{f,(o,r)} * o_o - (1 - b_{f,r}) * BigM \,\forall r \in TR, f \in F$ where $b_{f,r}$ is again the boolean relating to f for each period r, and $BigM$ the mathematical big M.

Run-time and SLA cost models

As presented in "Quality reduction of power flexibility" section, for the current model implementation quality reduction is viewed in terms of delay; for frequency scaling the delay is calculated with an adapted version of the model in Etinski et al. (2012), and for shifting, the delay is simply the number of shifted periods. The involved SLA costs are determined based on Garg et al. (2014) (see "Cost of power flexibility" section).

In the linear optimisation model, SLA costs are incorporated via the decision variable eSL_S that sums up all power consumption that has been shifted $s = r - o$ periods, either directly or indirectly by the modified computation time in the case of DVFS. The costs for shifting are linear to the amount of periods $s \in S$ the power consumption was shifted. eSL_S is multiplied by the costs for shifting per period CS_S to get the total SLA costs SLA.

$$eSL_S = \sum_{f \in F} \sum_{s \in S, (o,r) \in U : r - o = s} d_{f,(o,r)} * o_o \tag{13}$$

$$SLA = \sum_{s \in S} eSL_s * CS_s \tag{14}$$

Models for the power-flex markets

As mentioned, in this sample implementation of the optimisation framework, the three power flex markets EPEX, SCR and AGU are modelled and optimised against each other. Assuming that the data centre is a price taker on the market, historical price data can be used as given.

EPEX model

The EPEX market is modelled as a time-based vector with prices for each point in time. In accordance with "Optimisation objectives" section, the part of the objective function that maximises the net benefit of the EPEX market multiplies the energy consumption of each period with the difference between the baseline energy price and the pre-calculated EPEX price vector. The only constraint is the minimum trading requirement of 100 kWh (SE 2019). The participation fee is activated via a boolean variable if it turns out to be worth to engage in the EPEX market.

Secondary reserve market

The Secondary Control Reserve (SCR) model is composed of three parts: the costs of pre-qualification, the revenues for provisioning the load change (i.e. W, but calculated in Wh, see assumption 5), and the revenues for the actually 'delivered' electric energy (i.e. Wh) adaptation. As mentioned in Modelling power flexibility markets, the Secondary Control Reserve market includes four products m, all of which are taken into account.

The part of the objective function 11 that maximises the revenues from the participation in the Secondary Control Reserve market accounts for the two revenue streams, i.e. load price and energy price. The effort for the mandatory 'pre-qualification test' for market participation is interpreted as 'market entrance cost' C^{SCR} according to the terminology in "Modelling power flexibility markets" section; it is activated via a boolean and subtracted from the objective function.

In case the data centre is in the retrieval pool in week $w \in W$, where W is the weekly view on the optimisation time range TR, it has to deliver the load change service whenever activated by the network provider. The period length of the grid provider is a quarter of an hour; it is therefore assumed that the power flex offered in w, PF^{SCR}, is activated in the first 5-min period $aw \in AW$ and then needs to be maintained in the second and third 5-min periods $aw+s$. AW are the activation periods for the HPCC in which it needs to adapt its consumption. This is expressed by a set of constraints, of course only for the activation periods which are represented by a tuple including the activation week w and period aw within the week. Due to the assumption of certainty, these periods are known exactly.

Another constraint ensures that the DC not only meets the required amount of power adaptation $PF_w^{SCR,PosHT}$, but also does not over-fulfil this requirement; the German Transmission System Operator has set the limit to over-fulfillment oMA to 70%. Equation 15 shows this constraint in case the positive regulation service is activated in the high tariff time $PosHT$:

$$e_{aw} - e_{aw+s} \leq PF_w^{SCR,PosHT} * (1 + oMA),\ \forall (aw, w) \in PosHT, s = 1, 2. \tag{15}$$

The load price PP^{SCR} part of the revenue is paid if the data centre is chosen into the retrieval pool for the considered week. Adopting the procedure explained in "Modelling power flexibility markets" section leads to the information on the prices and of the

affected weeks. This means that for the affected weeks $w \in W$ load prices are multiplied with the offered load. For every product on the Secondary Control Reserve market, a decision variable is introduced, which expresses the offered load in this week's bidding process.

The energy-based part EP^{SCR} of the revenue is paid only if the DC's offered load is actually activated, indicated by the boolean variable b_w. Therefore the corresponding decision variables have the value of the offered load and are otherwise zero, as enforced by a set of constraints. To calculate the energy-based revenue, the activated energy is multiplied by two. The reason is that each SCR event time slot consists of 15 min; and, as explained before, the first DC 5-min slot is assumed to be spent on the adaptation, the next two 5-min slots are the compensated activation period. Each event can last up to one hour, which would extend the billing time. As the SCR data do not include the information if the activation lasted for the whole 15 min or if there were several subsequent activations, to not over-estimate the achieved market revenue, each 15 min period is treated separately. This means that in the case of a longer-lasting activation, the revenues in this model are slightly underestimated. Subsequently, this is multiplied by the pre-calculated energy price and added up over the weeks. All this is summarized in (16)–(19):

$$\max T^{SCR} - C^{SCR} \tag{16}$$

$$= \sum_{m} \left(T^{SCR,m} - 0.3 * T^{SCR,m} \right) \tag{17}$$

$$= \sum_{m} \left[\sum_{w} b_w * PF_w^m * EP_w^m * 2 + \sum_{w} PF_w^m * PP_w^m - 0.3 * T^{SCR,m} \right]; \tag{18}$$

$$m = \{PosHT, PosNT, NegHT, NegNT\}; aw \in AW, w \in W, \tag{19}$$

where T^{SCR} is the turnover on the Secondary Control Reserve market for all products $m = \{PosHT, PosNT, NegHT, NegNT\}$, the positive and negative reserve power, both in high and low tariff times.

Modelling Atypical Grid Usage

In the power control scheme Atypical Grid Usage, the regular power charge pc is not changed, however under this scheme it is not paid for the overall peak power $\hat{P}$ per year but the peak power in specific peak load windows $HT \in TR$.

The benefit resulting from the application of the Atypical Grid Usage strategy can be calculated by multiplying the difference between the current maximum load in all time windows, $\hat{P}$, and the one in the peak load time windows HT, $\hat{P}_{HT}$, with the load price pc. This leads to the Atypical Grid Usage part of the objective function:

$$\max \; T^{AGU} = \left(\hat{P} - \hat{P}_{HT} \right) * pc \tag{20}$$

$$\hat{P}_{HT} = \max_{ht} \left[\sum_{f} e_{ht,f} * pow_f * 5/60 \right]; \; ht \in HT; f \in F \tag{21}$$

where T^{AGU} is the turnover from participating in the Atypical Grid Usage scheme. Again, it should be noted, that due to the assumption of constant power in-between measurement points, the metric in the implementation is energy (Wh), not power (W) and therefore $*5/60$ is used to do the conversion.

As the concept of Atypical Grid Usage is targeted at big consumers it carries three constraints: materiality distance, bagatelle threshold, and materiality threshold.

***Materiality Distance*:** The materiality distance MD requires that there must be a significant difference between the maximum load in the peak load time windows $\hat{P}_{HT}$, and the maximum load in all other time periods in the year $\hat{P}$; this means: $\hat{P} - \hat{P}_{HT} \geq MD$

***Bagatelle Threshold*:** The bagatelle threshold BT requires the payment for the customer to be beyond a minimum BT to avoid system-wide losses; this means: $(\hat{P} - \hat{P}_{HT}) * pc \geq BT$

***Materiality Threshold*:** Finally, apart from the absolute materiality distance, the customer additionally will need to fulfil a significant revenue threshold. The following constraint reflects the materiality threshold MT, as given by the network providers: $\hat{P}_{HT}P \leq \hat{P} * (1 - MT)$

Integration of modelling components into complete model

In this section, the pieces of the optimisation objective presented in the preceding section will be put together. The glue is a simple addition of the gross benefit from the different markets of flexibility which basically can all be tackled at the same time. As long as there are no constraints like specific chunks of power flexibility delivered or an upper limit to offering power flexibility by one market participant, the algorithm chooses the most profitable market and offers the current flexibility to just this one. At each time step, a different market can be the most profitable one.

On the DC side, the model minimises the overall costs that are dependent on the technology of the power management strategies and their mix and the corresponding SLA cost. If the expected revenue is higher than the penalty cost, the model accepts penalties from SLA contracts.

The complete objective function is Eq. 11 which maximises the overall benefit as the difference between revenues and cost.

Results

The optimisation was carried out for different periods of time, chosen according to differences both on the power flex markets and the situation in the considered data centre. The first period, chosen as the basic run, comprises weeks 9-12 in March 2014, and the second period weeks 29-32 in July/August 2014. This run in March 2014 is the reference point for several sensitivity analysis runs which illuminate the dependency of the obtained results on the chosen parameters.

The subsequent sections will first present the parameters of the baseline scenario and the corresponding results, and then the various sensitivity runs.

Optimisation parameters

The parameters chosen for the baseline run are as close as possible to the considered scenario DC and markets:

- For the Sandy Bridge processors, nine specific frequencies between 1.2 and 2.7 can be chosen.
- We assume a flexibility for shifting of max 20% unless restricted by SLA cost.
- SLA costs are constructed using the approach described in "Cost of power flexibility" section. We calculate the costs/hour for the whole DC by using the number of nodes, the maximum load of the DC, and the costs of 0.36€/node hour (Stuttgart 2018). These are rather low, as results will show, but the most realistic choice for the DC at hand.
- As the baseline energy price of the considered data centre is not known, the influenceable part of the average industry price for Germany 2014 before taxes of 4.61 ct/kWh, was used (Bundesnetzagentur 2014).
- As the maximum adaptation for the SCR is smaller than the required SCR minimum adaptation size, it was assumed that the DC participates in SCR via an aggregator. For the aggregator, a fee of 30% of the SCR market turnaround was assumed. Prequalification cost is assumed to be covered by this fee[11].
- Regarding the Atypical Grid Usage, the bagatelle threshold was 500 €/year, materiality distance and threshold were 20% and 100 kW accordingly (Bundesnetzagentur 2011).
- The EPEX price vector used is the minimum of the EPEX Day Ahead auction price in every period[12].

Optimisation results

The optimisation run was executed on a 10 core CPU machine with 60 GB memory and completed after 42 min. The first period, 24th February - 23rd March 2014, was used as a baseline run. The second optimisation period, 14th July - 10th August 2014, was chosen mainly to check on the impact of atypical grid usage, which is not enacted in summer by the German Federal Network Agency. While the workload in the data centre in summer was slightly different from spring, its similarity in volume is another reason why these two time ranges were chosen for optimisation. The results can be found graphically in the Figs. 3 and 4; the values for some relevant metrics in the Tables 3 and 4.

Comparing the two figures already gives some hints at the nature of adaptation processes and additionally at the impact of workload and market characteristics. As can be seen, even though the volume of the workload is comparable in both months, it peaked more often in summer.

Analysing the workload on job level shows that on average run-time, power, and energy consumption per job was a little higher in summer. The number of nodes running per job also varied more in summer than in March 2014. Additionally, there was a noticeable workload drop at the end of the second optimisation phase. All of this explains the more fluctuating behaviour of the power profile of the data centre jobs in summer.

From the flex market side, the most obvious differences are the high density of SCR events in the third week of March which was followed by a price drop on the EPEX market. In summer there was no such disturbance. As mentioned, there are no atypical grid usage 'peak load windows' in summer.

Already, the optimisation results reveal the high impact of the flex market conditions on the behaviour of the resulting job power consumption. The impact of the Atypical

[11] This figure is based on a telephone call with the aggregator Next-Kraftwerke in June 2017
[12] http://www.epexspot.com/en/market-data/dayaheadauction/auction-table/2014-08-21/DE

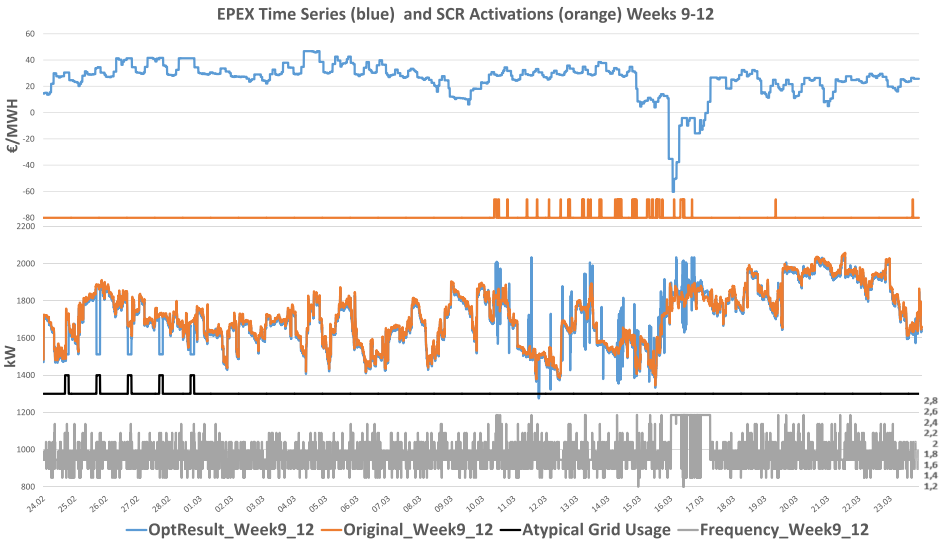

Fig. 3 Baseline optimisation run in March 2014

Grid Usage peak load windows is fully mirrored in the adaptation curve (see Fig. 3), and the seemingly erratic behaviour of the optimised job power curve caused by the adaption to the frequent SCR events in the third week of March is also conspicuous. The SCR events cause much less disturbance in summer, despite their being more numerous than in March. Table 3 sheds some light on the difference in adaptation processes between the baseline optimisation and the summer version. The highest income from using the power flex function of shifting and frequency scaling originates in both cases through engaging in the EPEX market (in both cases more than 99% of the total income). Even though there are more opportunities for profiting from SCR events in summer, these are hardly used (only 10% of the benefit realized from SCR in March); SLA costs, though neglectable, are nonetheless higher than in March. This leads to the conclusion that the EPEX market,

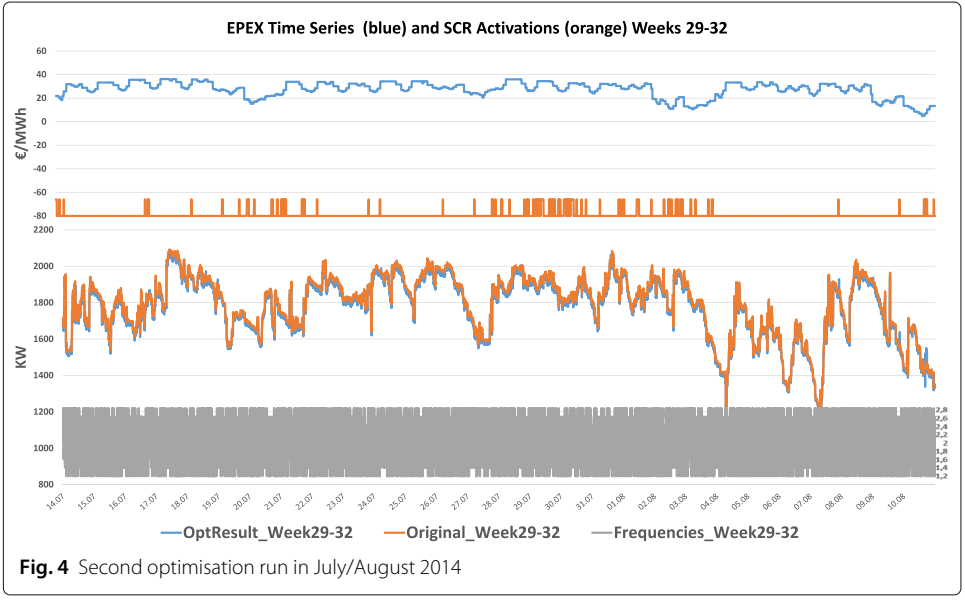

Fig. 4 Second optimisation run in July/August 2014

Table 3 Optimisation results: Economic impact

	March 2014		July/August 2014	
Items	Benefit	Cost	Benefit	Cost
EPEX	89,875		92,835	
Atypical grid usage	744		-	
SCR positive	555		80	
SCR negative	-		29	
SLA		340		7
Net Benefit	90,834		92,937	

in general, is considerably more profitable than the other markets. In summer, although the new maximum load (2.105 MW) was higher than the maximum load of the original time series, this has no impact on the benefit as the power fees are determined by the maximum load in the peak load time windows in March.

The high attractiveness of the EPEX market originates in the fact that, while the frequency in needed adaptations is high, only little adaptations are necessary for each time step. They can, therefore, be carried out by using frequency scaling. This strategy is less disruptive than shifting workload since allows for specific and efficient tuning of the power profile (see also formula in Production of power flexibility and Quality reduction of power flexibility). These conclusions are backed by the results of the optimisation algorithm with regards to the applied power flex strategies:

Although we assumed that 20% of the workload was flexible, only 0.21% of the load was actually shifted; and in the overwhelming number of cases, the workload was only shifted into the next period, and no longer. Figure 5 shows the result of a shifting event as a reaction to the high time window of atypical grid usage. The cost of shifting more load would have exceeded the achievable benefit, otherwise, the algorithms would have shifted the whole allowed range.

The power flex function of frequency changing was, in fact, responsible for most of the adaptation. As Table 4 shows, the average frequency changed from 2.40 to 1.87. The most efficient frequency, 1.8 GHz, was chosen in 67% of the periods, whereas without engaging in flexibility markets it was applied only in 11% periods. Since the frequency was in general higher and also more varied over time in the original July/August data, the applied frequency changes also took place more often due to a higher incentive to adapt and thereby generate energy savings.

Sensitivity analysis

To test the behaviour of the algorithm a set of sensitivity analyses was carried out with regards to

- the flexibility of the workload in terms of volume,

Table 4 Optimisation results: Impact on physical metrics

	March 2014		July/August 2014	
Items	Before Opt.	After Opt.	Before Opt.	After Opt.
Max load (MW)	1.658	2.034	1.658	2.105
Average frequency (GHz)	2.40	1.87	2.48	1.96
Tot. Energy Cons. (MWh)	1,158,962	1,148,215	1,203,789	1,190,404

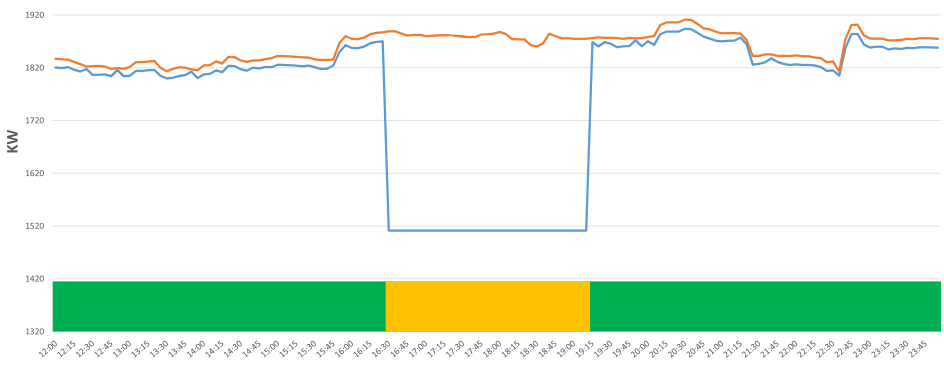

Fig. 5 Load Shifting in case of Atypical Grid Usage

- the variability of frequency,
- and the sensitivity to SLA cost.

All sensitivity runs were implemented for the time range in March specified as baseline optimisation. The results of these sensitivity runs will be presented in the following sections.

Flexibility of workload

As can be seen from the Table 5, being able to shift not only 20% of the workload, but 40% or even 80% increases the resulting income, but it also impacts the shares of the different flexibility markets that are used in the optimisation.

When comparing the optimisation results, one insight stands out. The values of the objective function increased almost linearly with the percentage of flexibility from 90,067 €in the baseline optimisation run, achieving net revenues of 368,630 €when 80% of the workload could be adjusted. One reason is that the assumed shifting cap limited the options to optimise on the price difference between the baseline price and the EPEX price. As this price difference was big, the higher the assumed potential of shifting, the higher the optimisation range. What is interesting, though, is that the mix of accessed power flex market changed. Just like Atypical Grid Usage, SCR is targeted at adapting higher amounts of power than the EPEX market is. With the higher share of power flexibility, power flexibility markets that relied on higher adaptive volumes were more enabled. This explains why the resulting revenues increased over-proportionally with the allowed workload flexibility. The share of workload shifting vs. frequency scaling therefore also increased, and as a consequence, this created SCR costs which behaved over-proportionally. Only in the case of an 80% workload flexibility, negative SCR was applied, i.e. that the power consumption of the data centre was increased in times where

Table 5 Sensitivity analysis: Workload flexibility in terms of volume

WL flex.	20% absolute	*20% percentage*	40% absolute	*40% percentage*	80% absolute	*80% percentage*
AGU	744	*0,8%*	2,666	*1,5%*	6,510	*1,8%*
EPEX	89,875	*99,8%*	180,641	*98,5%*	361,814	*98,2%*
SCR	555	*0,4%*	1,793	*0,5%*	2,731	*0,5%*
SLA cost	340	*0,6%*	994	*1%*	1,658	*0,7%*
Obj. value	90,067	*100%*	183,338	*100%*	368,630	*100%*

the overall power demand was too low compared to the energy supply, and reserve power thus became necessary.

Regarding the chosen frequencies, at a 40% flexible load, these exhibit almost the same distribution as in the baseline optimisation run. Of course, at a flexibility of 40%, the impact of the frequency adaptation was higher than if it was applied to a mere 20% of the load. With 80% flexible load, the most energy-efficient frequency 1.8 GHz was chosen in favour of all other frequencies at 22% points less compared to the baseline scenario of 20% flexible load share. In all three scenarios, the energy was only shifted up to two periods, corresponding to ten minutes.

Fixed frequency

The impact of a fixed frequency on the result of the objective value was significant. Fixed frequency means that the workload was computed using the original frequency, which for each time step was one average value of the load. This value could differ in each time step. If frequency scaling were to be turned off and only load shifting allowed, the benefit from the EPEX market would be reduced by a factor of more than 20, reaching only 4,293€. This again corroborates the observations in the baseline runs, that frequency scaling is more efficient than workload shifting.

As explained before, the SCR market load shifting is the preferred answer to SCR activation due to the higher amounts of required adaptive volumes. Up to 34% of the average load in the time range under consideration was offered on the SCR market in this sensitivity scenario, which is close to the maximum possible shifting volume of 20%.

Fulfilling the requirements of the SCR market was assumed to be mandatory in this model because once the bidding was successful the DC had to adjust if activated. Therefore, the requirements had to be fulfilled by using load shifting. This caused the SLAs to be twice as high as with DVFS, namely 685 €.

To fulfil the needs for SCR by shifting, an increase of the maximum load from previously 1658 kW to 1797 kW in peak load times was created which led to an increase in cost for network usage of 708 €. This implies that the option for AGU could not be exploited in this scenario. The results of this sensitivity run can be seen in the grey curve of Fig. 6. It follows more or less the baseline optimisation, however without all the little deviations originating in the EPEX adaption and with much higher amplitudes whenever SCR events occur.

Changing SLA cost

The sensitivity of the optimisation to SLA costs was examined in two runs without any SLA costs. For this scenario, artificial constraints for shifting the load had to be implemented since the linear optimisation model requires a defined optimisation range (see "Power flexibility through workload shifting" section). In the first run, the median job duration was used as a flexibility range constraint, which was best reflected by two periods, corresponding to ten minutes. This means it was only allowed to shift jobs by the time range of the median job duration, thus more than doubling the execution time of half of the jobs. The second run used the average job duration, which was best represented using 43 periods, i.e. roughly 3.5 h. Compared to the basic scenario, the total net benefit

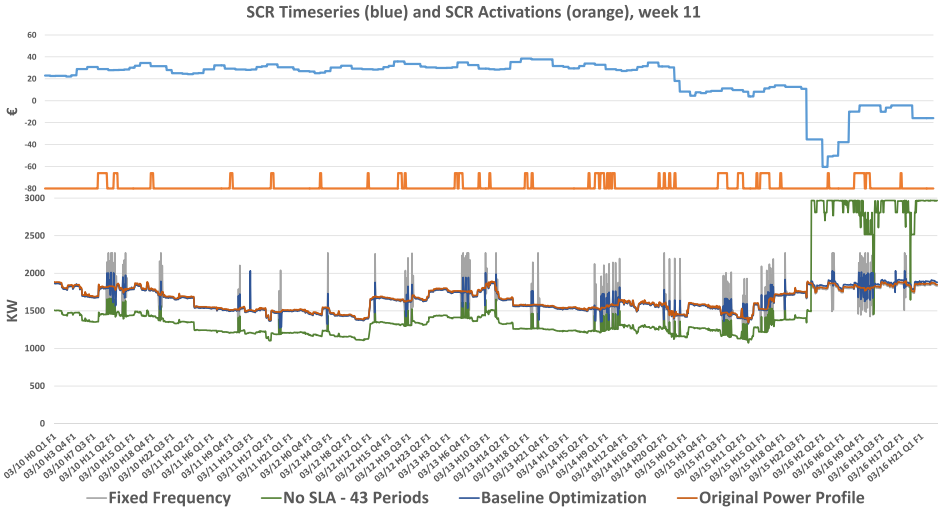

Fig. 6 Two Sensitivity Runs: Fixed Frequency & Lack of SLA Cost

increased by roughly 6000 €to 95,701 €in the first case. In the second case, i.e. extending the flexibility range by a factor of more than 20, the total objective value increased by about 7000 €.

In the first case, even though shifting costed nothing, 88% of the workload remained unshifted. Only 10% was shifted by one period and as little as 2% was shifted by two periods. In the second scenario quite a large share, 86%, was also not shifted at all, and the rest was distributed almost evenly to all periods where it was allowed to be shifted to. As in the other optimisation runs, the EPEX market was the most profitable power flexibility market. As can be seen from the green curve in Fig. 6, the adaptation seems to be just scaled-down compared to the baseline optimisation, until the very last day of the presented week (March 16th). At this point, both SCRs offered some longer-lasting opportunities and the EPEX prices were considerably decreasing.

The results of changing the cost parameters support the overall finding, that the benefit of abolishing SLA costs does not increase in proportion with the flexibility range. On the other hand, it shows how small variations might lead to a comparably high impact. For instance, being allowed to shift jobs by merely 10 min increased the overall benefit by more than 6%.

Discussion and outlook

In times of an energy transition from fossil fuels to renewable energies, there is a rising and necessary focus on power consumers who can adapt their demands.

This paper introduced a framework model for DR with DC which is not limited to assessing the technical DR potential but takes the economic issue into account and thus analyses the economic potential. It even offers some starting points that go beyond the economic potential, albeit those were not evaluated. The approach was evaluated through an optimisation modelling instance on three flexibility markets, applying two power flex production functions. The implemented version of the modelling framework shows that on German power flex markets a multiple-market strategy is significantly beneficial. Price-based incentives are extremely effective in reducing the costs of energy purchase,

yet the DC is not compelled to adapt its load curve. Furthermore, HPC DCs with a mature demand side management could offer their flexibility on the SCR markets, however in doing so would then be bound to adapting their load curve in the event of an activation. Participating in the reserve markets may necessitate higher levels of automation in the communication and implementation of DR requests due to timing requirement. Automation has therefore been identified as 'technical enabler' of DR from very early on. The first quasi-standard 'OpenADR' was developed by LBNL in 2010 and frequently applied mostly in the U.S. in the subsequent years; SEP2.0 and Green Button are less well-known open standards (Shoreh et al. 2016). Meanwhile an international standardization organization, the International Electrotechnical Commission (IEC) has taken the baton and has been working on a set of standards in the context of the digitization of the energy systems including DR (Dong et al. 2017). The specifically targeted IEC 62746 series standard has just been released[13]. Also, China has acknowledged the challenges of the next generation power grid and is working on its standards (Dong et al. 2017).

Atypical Grid Usage, it must be concluded, is the concept with the least risk associated. The peak load times for the year are published a long time in advance, meaning that the cost savings can be calculated with precision. However, the by far most beneficial power flex market is the EPEX market where small but steady adaptations lead to high revenue. This is especially true in the combination of DVFS with Load Shifting, which allows for flexibility without the need to shift large volumes of the load. The sensitivity analysis showed that not using DVFS leads to much lower revenues. Besides that, it found that the relation of the share of the load which is flexible with the turnovers is almost linear.

While this paper shows that it is of high worth for DCs to branch into flexibility markets, more detailed models (probably via simulation) of DR with a HPCC on a more fine-grained job level would lead to more precise results. Also entering uncertainty into the model on both the DC and the power flex market side would add to the reliability of results. This would provide the possibility for Data Centres to build on these models and apply them in the operational business. Detailed simulations of the markets, HPCCs, and their interaction would provide valuable insights to ease the integration of these power consumers. Building on this framework, more power flex functions should be incorporated and validated in the future. Taking into account the risk adversity of High Performance Computing Centres could help clarify why their use for Demand Response is not widespread in practice. This is part of our planned future research.

These results are but one potential evaluation of the suggested modelling framework for data centre demand response. Other application options are by positioning new findings in terms of which DR potential is targeted, which power management strategies employed and to which extent interdependencies are accounted for. It is a major strong point of this model that it can be applied to other types of DCs, involving other sets of power management strategies as well as other power flex markets. The main result of this paper is, therefore, a generic modelling framework focusing on the power flexibility of DCs and thus offering a backbone to understanding and evaluating data centre demand response.

Acknowledgements
We are grateful for valuable discussions and feedback from Prof. Christian Becker, University of Mannheim as well as from Martijn Schootuiterkamp and Dr Marco Gerards from the University of Twente.

[13] https://webstore.iec.ch/publication/26267

Authors' contributions
Sonja Klingert has focused on the framework and has been guiding the work on the linear optimisation problem; Sebastian Szilvas has implemented the linear optimisation problem. The author(s) read and approved the final manuscript.

Funding
The authors were not externally funded.

Availability of data and materials
The datasets regarding the data centre generated and analysed during the current study are not publicly available due to restrictions of the provider of the data centre but are available from the corresponding author on reasonable request. The sources of the publicly available datasets are stated as references within the paper.

Competing interests
The authors declare that they have no competing interests.

Received: 22 May 2020 Accepted: 13 July 2020
Published online: 03 August 2020

References

Aksanli B, Rosing T (2014) Providing regulation services and managing data center peak power budgets. In: Proceedings of the Conference on Design, Automation & Test in Europe, IEEE. p 143. European Design and Automation Association
Auweter A, Bode A, Brehm M, Brochard L, Hammer N, Huber H, Panda R, Thomas F, Wilde T (2014) A case study of energy aware scheduling on supermuc. In: International Supercomputing Conference. Springer. pp 394–409. https://doi.org/10.1007/978-3-319-07518-1_25
Barth L, Ludwig N, Mengelkamp E, Staudt P (2018) A comprehensive modelling framework for demand side flexibility in smart grids. Comput Sci-Res Dev 33(1-2):13–23
Basmadjian R, Botero JF, Giuliani G, Serra XH, Klingert S, De Meer H (2016) Making data centres fit for demand response: Introducing greensda and greensla contracts. IEEE Trans Smart Grid 9(4):3453–3464
Berl A, Klingert S, Beck MT, de Meer H (2013) Integrating data centres into demand-response management: a local case study. In: Industrial Electronics Society, IECON 2013-39th Annual Conference of the IEEE. IEEE. pp 4762–4767
Bundesnetzagentur (2011) Leitfaden zur Genehmigung Von Individuellen Netzentgelten Nach §19 Abs. 2 S. 1 StromNEV und Von Befreiungen Von Den Netzentgelten Nach §19 Abs. 2 S. 2 StromNEV. https://www.bonn-netz.de/Stromnetz/Preisblaetter/Preisblaetter/201109-Leitfaden-19-StromNEV-NetzA.pdf. Accessed 30 Apr 2020
Bundesnetzagentur (2014) Monitoringbericht 2014. Technical report, Bundesnetzagentur. S. 154. https://www.bundesnetzagentur.de/SharedDocs/Downloads/DE/Allgemeines/Bundesnetzagentur/Publikationen/Berichte/2014/Monitoringbericht_2014_BF.pdf?__blob=publicationFile&v=3
Bundesverband der Energie- und Wasserwirtschaft (2019) Nettostromverbrauch in Deutschland in den Jahren 1991 bis 2018. Technical report, BDEW
Chamberlin JH, Gellings CW (1988) Demand-Side Management: Concepts and Methods. Fairmout Press Inc, Lilburn
Chen H, Caramanis MC, Coskun AK (2014) Reducing the data center electricity costs through participation in smart grid programs. In: Green Computing Conference (IGCC), 2014 International. IEEE. pp 1–10
Cioara T, Anghel I, Bertoncini M, Salomie I, Arnone D, Mammina M, Velivassaki T-H, Antal M (2016) Optimized flexibility management enacting data centres participation in smart demand response programs. Futur Gener Comput Syst. https://doi.org/10.1016/j.future.2016.05.010
Coalition SED (2017) Explicit demand response in europe - mapping the markets 2017. Technical report, Technical report, Brussels
Commission E (2013) Incorporating demand side flexibility, in particular demand response, in electricity markets: Delivering the internal electricity market and making the most of public intervention. Communication from the Commission, SWD (2013) 442 final
Cupelli LJ, Schutz T, Jahangiri P, Fuchs M, Monti A, Muller D (2018) Data center control strategy for participation in demand response programs. IEEE Trans Ind Inf. https://doi.org/10.1109/tii.2018.2806889
der Justiz und Verbraucherschutz B (2018) Stromnetzentgeltverordnung §19 Abs. 2. https://www.gesetze-im-internet.de/stromnev/__19.html. Online; Accessed 18 April 2018
Dong M, Tian S, Zhu W, Jia B, Li B, Qi B (2017) Research and development of automated demand response standard system. In: 2017 2nd International Conference on Power and Renewable Energy (ICPRE). IEEE. pp 608–611. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8390607
Elnozahy EN, Kistler M, Rajamony R (2002) Energy-efficient server clusters. In: PACS, vol. 2325. Springer. pp 179–196. https://doi.org/10.1007/3-540-36612-1_12
Energiewende A (2019) European Energy Transition 2030: The Big Picture. https://www.agora-energiewende.de/fileadmin2/Projekte/2019/EU_Big_Picture/153_EU-Big-Pic_WEB.pdf. Online; Accessed 28 Jun 2020
Energiewende A (2020) The European Power Sector in 2019. https://www.agora-energiewende.de/fileadmin2/Projekte/2019/Jahresauswertung_EU_2019/172_A-EW_EU-Annual-Report-2019_Web.pdf. Online; Accessed 28 Jun 2020
entsoe (2019a) Manually Activated Reserves Initiative. https://www.entsoe.eu/network_codes/eb/mari. Online; Accessed 3 July 2020
entsoe (2019b) PICASSO. https://www.entsoe.eu/network_codes/eb/picasso. Online; Accessed 3 July 2020
entsoe (2019c) TERRE. https://www.entsoe.eu/network_codes/eb/terre. Online; Accessed 3 July 2020
Etinski M, Corbalán J, Labarta J, Valero M (2012) Understanding the future of energy-performance trade-off via dvfs in hpc environments. J Parallel Distrib Comput 72(4):579–590

(EU) CR (2017) Commission Regulation (EU) 2017/2195 of 23 November 2017 establishing a guideline on electricity balancing (Text with EEA relevance.) C/2017/7774. https://eur-lex.europa.eu/eli/reg/2017/2195/oj. Online; Accessed 28 Jun 2020
Fernández-Montes A, Fernández-Cerero D, González-Abril L, Álvarez-García JA, Ortega JA (2015) Energy wasting at internet data centers due to fear. Pattern Recogn Lett 67:59–65
Fleten S, Wallace SW, Ziemba WT (1997) Portfolio management in a deregulated hydropower based electricity market. Hydropower'97: Proc 3rd Int Conf Hydropower Trondheim/Norway/30 June-2 July, 1997. 97(1):197–204
Fleten S-E, Wallace SW, Ziemba W (2002) Hedging electricity portfolios via stochastic programming. Decis Making Under Uncertain 128:71–93. https://doi.org/10.1007/978-1-4684-9256-9
Fridgen G, Keller R, Thimmel M, Wederhake L (2017) Shifting load through space–the economics of spatial demand side management using distributed data centers. Energy Policy 109:400–413
Garg SK, Toosi AN, Gopalaiyengar SK, Buyya R (2014) Sla-based virtual machine management for heterogeneous workloads in a cloud datacenter. J Netw Comput Appl 45:108–120
Ghamkhari M, Mohsenian-Rad H (2012) Data centers to offer ancillary services. In: Smart Grid Communications (SmartGridComm), 2012 IEEE Third International Conference On. IEEE. pp 436–441. https://doi.org/10.1109/smartgridcomm.2012.6486023
Ghatikar G (2012) Demand response opportunities and enabling technologies for data centers: Findings from field studies
Ghatikar G, Piette MA, Fujita S, McKane A, Dudley JH, Radspieler A, Mares K, Shroyer D (2009) Demand response and open automated demand response opportunities for data centers. Technical report, Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States)
Giacobbe M, Celesti A, Fazio M, Villari M, Puliafito A (2015) Towards energy management in cloud federation: a survey in the perspective of future sustainable and cost-saving strategies. Comput Netw 91:438–452
Gils HC (2014) Assessment of the theoretical demand response potential in europe. Energy 67:1–18
Glanz J (2012) Power, pollution and the internet. The New York Times 22
GmbH NK (2018) Best of 96. https://www.next-kraftwerke.de/virtuelles-kraftwerk/stromverbraucher/variabler-stromtarif. Online; Accessed 20 Feb 2018
Haubrich H, Zimmer C, Sengbusch KV, Li F (2001) Analysis of electricity network capacities and identification of congestion. Technical report, Institut für elektrische Anlagen und Energiewirtschaft. http://www.iaew.rwth-aachen.de/publikationen/EC_congestion_final_report_appendix.pdf
Hintemann R (2018) Digitalisierung treibt Strombedarf von Rechenzentren: Boom führt zu deutlich steigendem Energiebedarf der Rechenzentren in Deutschland im Jahr 2017. Technical report, Borderstep Institute for Innovation and Sustainability
Islam MA, Ren X, Ren S, Wierman A, Wang X (2016) A market approach for handling power emergencies in multi-tenant data center. In: High Performance Computer Architecture (HPCA), 2016 IEEE International Symposium On. IEEE. pp 432–443. https://doi.org/10.1109/hpca.2016.7446084
Jochem E, Adegbulugbe A, Aebischer B, Bhattacharjee S, Gritsevich I, Jannuzzi G, Jaszay T, Baran Saha B, Worrell E, Fengqi Z, et al (2000) Energy end-use efficiency. Technical report, UNDP/UNDESA/WEC: Energy and the Challenge of Sustainability. World Energy
Kirpes B, Klingert S (2016) Evaluation process of demand response compensation models for data centers. In: Proceedings of the 5th International Workshop on Energy Efficient Data Centres. ACM. p 4. https://doi.org/10.1145/2940679.2940683
Klingert S (2018) Mapping data centre business types with power management strategies to identify demand response candidates. In: Proceedings of the Ninth International Conference on Future Energy Systems. ACM. pp 492–498
Klingert S, Becker C (2017) Economics-inspired modeling of data centre power flexibility. Comput Sci-Res Dev 33.1-2(2018):247–249
Kong F, Liu X (2015) A survey on green-energy-aware power management for datacenters. ACM Comput Surv (CSUR) 47(2):30
Liu Z, Chen Y, Bash C, Wierman A, Gmach D, Wang Z, Marwah M, Hyser C (2012) Renewable and cooling aware workload management for sustainable data centers. In: ACM SIGMETRICS Performance Evaluation Review, vol. 40. ACM. pp 175–186. https://doi.org/10.1145/2254756.2254779
Liu Z, Lin M, Wierman A, Low SH, Andrew LL (2011) Greening geographical load balancing. In: Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems. ACM. pp 233–244. https://doi.org/10.1145/1993744.1993767
Liu Z, Wierman A, Chen Y, Razon B, Chen N (2013) Data center demand response: Avoiding the coincident peak via workload shifting and local generation. Perform Eval 70(10):770–791
Mahmud AH, Ren S (2013) Online capacity provisioning for carbon-neutral data center with demand-responsive electricity prices. ACM SIGMETRICS Perform Eval Rev 41(2):26–37
Mas-Colell A, Whinston MD, Green JR (1995) Microeconomic Theory. Oxford University press, New York
Microsoft (2018) Azure Cloud SLAs Batch Jobs. https://azure.microsoft.com/de-de/support/legal/sla/batch/v1_1/. Online; Accessed 11 May 2018
Niedermeier F, Kazhamiaka F, de Meer H (2016) Energy supply aware power planning for flexible loads. In: Proceedings of the 5th International Workshop on Energy Efficient Data Centres. ACM. p 2
Patki T, Bates N, Ghatikar G, Clausen A, Klingert S, Abdulla G, Sheikhalishahi M (2016) Supercomputing centers and electricity service providers: a geographically distributed perspective on demand management in europe and the united states. In: International Conference on High Performance Computing. Springer. pp 243–260. https://doi.org/10.1007/978-3-319-41321-1_13
Piette MA, Watson D, Motegi N, Kiliccote S, Xu P (2006) Automated critical peak pricing field tests: Program description and results. https://doi.org/10.2172/901672
Qureshi A, Weber R, Balakrishnan H, Guttag J, Maggs B (2009) Cutting the electric bill for internet-scale systems. In: ACM SIGCOMM Computer Communication Review, vol. 39. ACM. pp 123–134

Rajaraman R, Kirsch L, Alvarado FL, Clark C (2002) Optimal Self-Commitment Under Uncertain Energy and Reserve Prices. Next Gener Electric Power Unit Commitment Models 36(April):93–116. https://doi.org/10.1007/0-306-47663-0_6. Accessed 30 Apr 2020
Rao L, Liu X, Xie L, Liu W (2010) Minimizing electricity cost: Optimization of distributed internet data centers in a multi-electricity-market environment. Proc IEEE INFOCOM. https://doi.org/10.1109/INFCOM.2010.5461933
ScaleMatrix (2018) ScaleMatrix SLAs. https://www.scalematrix.com/ntt/sla. Online; Accessed 11 May 2018
SE ES (2019) Day-ahead auction with delivery on the German/Luxembourg TSO zones. http://www.epexspot.com/en/product-info/auction/germany-luxembourg. Online; Accessed 17 Nov 2019
Services AW (2018) Amazon Web Services EC2 SLAs. https://d1.awsstatic.com/legal/amazon-ec2-sla/Amazon_EC2_Service_Level_Agreement_de.pdf. Online; Accessed 11 May 2018
Shoreh MH, Siano P, Shafie-khah M, Loia V, Catalão JP (2016) A survey of industrial applications of demand response. Electr Power Syst Res 141:31–49
Shoukourian H, Wilde T, Auweter A, Bode A (2015a) Power variation aware configuration adviser for scalable hpc schedulers. In: High Performance Computing & Simulation (HPCS), 2015 International Conference On. IEEE. pp 71–79. https://doi.org/10.1109/hpcsim.2015.7237023
Shoukourian H, Wilde T, Auweter A, Bode A, Tafani D (2015b) Predicting energy consumption relevant indicators of strong scaling hpc applications for different compute resource configurations. In: Proceedings of the Symposium on High Performance Computing, ACM. pp 115–126. Society for Computer Simulation International
Šikšnys L, Valsomatzis E, Hose K, Pedersen TB (2015) Aggregating and disaggregating flexibility objects. IEEE Trans Knowl Data Eng 27(11):2893–2906
Stewart GL, Koenig GA, Liu J, Clausen A, Klingert S, Bates N (2019) Grid accommodation of dynamic HPC demand. In: Proceedings of the 48th International Conference on Parallel Processing: Workshops. pp 1–4. https://doi.org/10.1145/3339186.3339214
Stuttgart H (2018) Lesefassung der HLRS Entgeltordnung 2017. https://www.hlrs.de/fileadmin/sys/public/solution_services/system_access/legal_requirements/HLRS_Entgeltordnung_Lesefassung_2017.pdf. Online; Accessed 9 Sept 2019
Tang C-J, Dai M-R, Chuang C-C (2013) Exploring the potential benefits of demand response in small data centers. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, IMECS
van der Veen RA, Hakvoort RA (2016) The electricity balancing market: Exploring the design challenge. Util Policy 43:186–194
Ventosa M, Baíllo Á, Ramos A, Rivier M (2005) Electricity market modeling trends. Energy Policy 33(7):897–913. https://doi.org/10.1016/j.enpol.2003.10.013
Virtual Iron Software I (2007) The new economics of virtualization. Technical report, Virtual Iron Software, Inc
Wang R, Kandasamy N, Nwankpa C (2012) Data centers as demand response resources in the electricity market: Some preliminary results. In: Intl. Workshop on Feedback Computing, USENIX
Wang R, Kandasamy N, Nwankpa C, Kaeli DR (2013) Datacenters as controllable load resources in the electricity market. In: 2013 IEEE 33rd International Conference on Distributed Computing Systems. IEEE. pp 176–185. https://doi.org/10.1109/ICDCS.2013.16
Wang C, Urgaonkar B, Wang Q, Kesidis G (2014) A hierarchical demand response framework for data center power cost optimization under real-world electricity pricing. In: Modelling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2014 IEEE 22nd International Symposium On. IEEE. pp 305–314. https://doi.org/10.1109/mascots.2014.45
Whitney J, Delforge P, Masanet E, Peridas G, Stamas M, Jimenez N, Remick P, Clinger J, Brown SMT (2014) Scaling up energy efficiency across the data center industry: Evaluating key drivers and barriers. Issue Paper No. IP:14–08. Natural Resources Defense Council (NRDC)
Wierman A, Liu Z, Liu I, Mohsenian-Rad H (2014) Opportunities and challenges for data center demand response. In: Green Computing Conference (IGCC), 2014 International. IEEE. pp 1–10. https://doi.org/10.1109/igcc.2014.7039172
Xu H, Li B (2014) Reducing electricity demand charge for data centers with partial execution. In: Proceedings of the 5th International Conference on Future Energy Systems. ACM. pp 51–61
Yao J, Liu X, Zhang C (2014) Predictive electricity cost minimization through energy buffering in data centers. IEEE Trans Smart Grid 5(1):230–238
Zhou Q, Bialek JW (2005) Approximate model of European interconnected system as a benchmark system to study effects of cross-border trades. IEEE Trans Power Syst 20(2):782–788. https://doi.org/10.1109/TPWRS.2005.846178

Publisher's Note