Towards Integrating Data Centers and Energy Grid: A Theoretical Study on Geographic Load Shifting

Carina Baur*, Theresa Paulus†, Frank Loh*, Krzysztof Rudion†, Tobias Hoßfeld*
*University of Würzburg, Chair of Communication Networks, Germany
† University of Stuttgart, Institute of Power Transmission and High Voltage Technology, Germany
Contact: carina.baur@uni-wuerzburg.de

***Abstract*—The increasing geographical decoupling of electricity generation and consumption, driven by the growth of decentralized renewable energy, poses challenges for grid stability and transmission. These challenges lead to additional costs for grid operators due to inter-operator payments and redispatch, and with growing decentralization of power generation, small-scale loads and flexibility resources are intended to be utilized for redispatch to relieve the energy grid. Distributed data centers offer an opportunity to provide demand-side flexibility due to the spatial adaptability of their workloads. This paper investigates their potential to support the German electricity grid by shifting workloads toward regions with renewable energy availability. Using real-world generation data and realistic workload and data center assumptions, we model spatial workload shifting with regards to available renewable energy across Germany as an optimization problem. We investigate the flexibilities in power draw, associated CO_2 savings, and performance trade-offs under three geographic scenarios and two operational strategies. Results indicate that spatial load shifting can enhance renewable energy utilization, reduce CO_2 emissions, and provide meaningful flexibilities for the energy grid. The findings also highlight the limiting factors, such as data center capacity and temporal workload availability.**

***Index Terms*—Geographic load shifting, Power flexibilities of data centers, CO_2 emission reduction**

I. Introduction

Electricity generation and consumption are increasingly decentralized. Renewable energy sources such as wind and solar are often situated far from demand centers, placing significant stress on transmission infrastructure and driving up redispatch costs. Moreover, the variability of renewable generation creates challenges for maintaining grid stability. Surplus renewable energy is frequently curtailed while fossil-based generation remains active elsewhere due to transmission bottlenecks and despite renewable priority feed-in [1], [2].

On the demand side, large-scale data centers account for a rapidly growing share of electricity demand and are typically located near population centers and network hubs rather than renewable-rich regions. According to the German Federal Ministry for Economic Affairs and Climate Action (BMWK), German data centers currently consume around 20 TWh yearly, i.e., around 4 % of the total German electricity consumption, with projections of up to 80 TWh by 2045 [3], [4]. An overview of studies examining future data center energy demand covering individual countries, global estimates, and AI-focused analyses is provided in [5]. Unlike traditional loads, however, data center workloads exhibit spatial and temporal flexibility. Computational tasks can often be shifted between geographically distributed facilities providing demand-side flexibility. Exploiting this flexibility allows data centers to align their electricity consumption with renewable generation, potentially reducing CO_2 emissions and alleviating grid congestion, as highlighted by the BMWK report [3]. However, shifting load increases latency, as data has to travel through the network, and consequently degrade service quality.

ISBN 978-3-903176-82-9 © 2026 IFIP

Beyond environmental benefits, demand-side flexibility is important from market and regulatory perspectives. Transmission system operators (TSOs) rely on flexible consumption to manage congestion and integrate high shares of renewables, with policy initiatives targeting smaller and distributed loads [6], [7]. In this context, data centers emerge as promising participants in future flexibility markets, where even comparatively small loads may be compensated for providing grid-support [8]. This creates an opportunity for operators to optimize performance, costs, and sustainability simultaneously.

In this work, we investigate the potential of geographically distributed data centers to support the German energy system by moving computational load towards regions with renewable energy surpluses. Using real-world electricity generation data and realistic workload assumptions, we model the achievable flexibility, associated CO_2 savings, and performance trade-offs under different operational strategies. By linking data center operation to regional renewable availability, this paper aims to bridge the gap between availability and demands in the energy grid and communication infrastructure. Limitations include that the reported savings are purely theoretical. The presented results reflect best-case scenarios and therefore represent upper bounds of the achievable emission reductions.

Therefore, the contribution of this work is three-fold: First, we present a modeling approach that utilizes geographically distributed data center workloads, regionally resolved renewable generation and electrical load, and CO_2 intensity in the German electricity grid, using real-world data from transmission system operators and Internet traffic measurements. Second, we formulate spatial workload shifting as an optimization problem that prioritizes renewable energy utilization. Third, we quantify the power flexibility of data centers under different operational strategies, investigating worst and best-case scenarios, and analyze potential effects on CO_2 reductions.

Consequently, this work is guided by the following three research questions (RQs):

- **RQ1:** Which flexibility potentials can geographically distributed data centers provide for the German electricity grid and which factors limit them?
- **RQ2:** When is workload shifting in geographically distributed data centers most effective in leveraging renewable energy availability?
- **RQ3:** To what extent can geographically distributed data centers reduce CO_2 emissions by shifting computational load toward regions with renewable energy surpluses?

In the remainder, Section II reviews background and related work. Section III describes the data, and Section IV details the methodology, including the system model, optimization problem, scenarios, and metrics. Section V presents the results, and Section VI discusses applicability and concludes.

II. Background

This section introduces background, beginning with the data center and networking infrastructure landscape. Then, information on renewable production and energy grid operation is given. It concludes with related work.

A. Data Centers and Network Infrastructure

Modern data centers are large-scale facilities that host compute, storage, and networking resources to deliver digital services. Their IT infrastructure comprises servers that execute workloads and network components such as switches and routers that interconnect these servers. Beyond that, data centers also include supporting infrastructure such as cooling [5]. Traffic between geographically distributed sites is carried over optical fiber infrastructure with high throughput.

In contrast to this network-level performance, application-level performance requirements determine how sensitive workloads are to network delay. Interactive and latency-critical services (e.g., real-time control or user-facing transactions) often require end-to-end latencies below a few tens of milliseconds [9], whereas batch processing, analytics, and many machine learning tasks are delay-tolerant and can be relocated with limited Quality of Service impact.

B. Renewable Generation

Modern electricity systems are undergoing a transition towards high shares of renewable generation, particularly wind and solar [10]. Their weather-dependent output and natural resource-dependent location introduces significant temporal and spatial variability, which increases the complexity of balancing generation and demand. For example in Germany, renewable feed-in already supplies more than half of electricity demand annually [11], with strong geographic asymmetries: wind power dominates in the northern states with offshore plants, whereas solar plants contribute more in the south [1]. A key metric to describe the temporal imbalance is residual load, defined as the difference between total demand and renewable generation. Comparing the residual load across different regions also provides insight into spatial imbalances. Regions

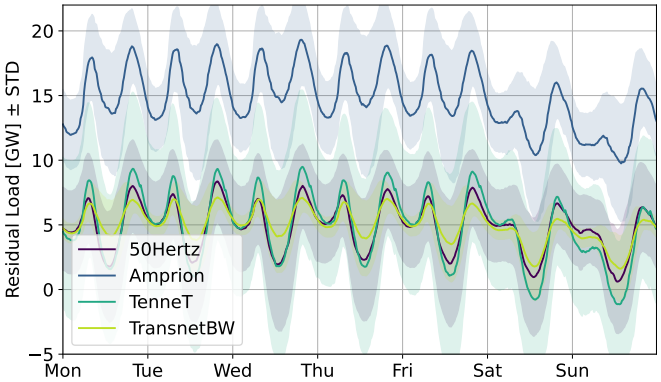

Figure 1: Mean residual load during the week for 2025. Shaded area shows standard deviation.

with negative residual load exhibit renewable surpluses and often face curtailment, while regions with high positive residual load depend heavily on conventional generators, resulting in higher CO_2 intensities [2]. Figure 1 visualizes the average residual load in gigawatts during a week in 2025 with data provided from the SMARD platform operated by the German Federal Network Agency (Bundesnetzagentur) [12]. Shaded areas show the standard deviation. The data is resolved for the four German TSOs in color, which cover different geographic zones. Clear daily and weekly patterns are visible, typical for renewable production and load in the energy grid. The residual load drops during the night due to less energy demand. It also drops around noon, as solar power generation increases during this time. During the weekends residual load is generally lower as demand is reduced. This leads to a mean negative load around noon on weekends in the TenneT and 50Hertz zones in northern Germany. Meanwhile, Amprion's residual load is comparatively high, as it covers North Rhine-Westphalia, the federal state with the highest population and a legacy of coal-powered energy production.

C. Congestion Management in the Energy Grid

With the increasing decentral production of renewable energy, many countries face challenges in grid congestion. The Netherlands is a prominent example where the rapid expansion of renewable generation has outpaced grid expansion and led to significant grid congestion, now being the limiting factor for further renewable deployment [13]. In Germany, due to the geographic imbalance between renewable generation dominated by wind in the north and major demand centers in the west and south [3], north–south transmission corridors are frequently congested [14], posing challenges for grid stability and increasing system costs.

To maintain grid stability during renewable availability or high load scenarios, system operators rely on flexibilities including adjustable generation that can shift in time and location [13]. This redispatchment mitigates congestion in constrained grid segments. Additional costs occur due to compensation for curtailed production and for additional generation in deficit regions. These costs more than tripled between

2020 and 2022 in Germany, exceeding 4 billion euros [13]. Traditionally, large industrial producers have provided these redispatch services. However, with the increasing decentralization of the energy grid, both regulatory frameworks and market designs increasingly encourage smaller, highly distributed flexible producers such as home PV installments, but also consumers, integrating previously untapped potential. In Germany, these efforts are known as Redispatch 2.0, and the following Redispatch 3.0 [6], [7], [8]. This presents an opportunity for data center operators: by shifting computational load toward renewable-rich regions, they can provide spatial flexibility and help to mitigate congested links. In return, they may receive monetary compensation, allowing environmentally beneficial load shifting to become an economic strategy.

D. Related Work

Rahman et al. [15] provide a survey of geographic load balancing (GLB), classifying objectives such as minimizing electricity cost, reducing carbon footprint, and maximizing renewable energy usage. Schiller et al. [16] formulate spatial load shifting as an optimization problem with regards to cost, CO_2 emissions, and water consumption, across distributed data centers in Europe. Hu et al. [17] extend on this by coupling distributed data center operation with smart grid dynamics.

Beyond price-driven GLB, several works exploit geographical flexibility to better align computation with renewable energy availability. Li et al. [18] argue for redesigning data center architectures and workload management to absorb renewable variability through spatial diversity and workload migration, reducing dependence on grid power. Liu et al. [19] show that geographical load balancing can be leveraged not only for cost reduction but also to reduce conventional energy consumption, demonstrating that under appropriate dynamic pricing schemes tied to carbon intensity, GLB can significantly increase the use of renewable energy. Similarly, Camus et al. [20] study cooperative self-consumption in geo-distributed clouds powered by on-site photovoltaics. Additional to these geo-distributed approaches, Agarwal et al. [21] focus on managing server clusters supplied by renewable energy sources. Complementary, Diamanti et al. [22] study resource allocation in multi-layer edge–fog systems using contract and game theory. Their work illustrates how energy-aware coordination can be achieved, especially under incomplete information through economic mechanisms, which translate to the implementation of geographic load shifting between data centers.

Geographic load-distribution as a tool to optimize data center operation with regards to different goal metrics is widely studied in literature, as well as economic incentives for the practical implementation. Beyond that, recent work considers energy-aware placement in edge–cloud and serverless environments, where latency, resource utilization, and energy consumption are explicit key performance indicators [23], [24]. These studies show that workload placement strongly affects both delay and energy efficiency, underscoring energy as a central system-level design metric.

However, our contribution lies in integrating these computational flexibilities with energy grid dynamics. We evaluate geographic load shifting not only as a data center optimization problem, but as a mechanism to relieve energy grid stress and congestion, thereby bridging the gap between ICT-level optimization and power system operation.

III. Datasets

This section presents the energy and computational workload data input required for our model.

A. Energy Grid Data

To analyze different spatial resolution scenarios two complementary datasets are used. The objective is to derive a time series of the residual load expressed as mean hourly power values for the selected spatial resolution. For the first scenario, which considers a spatial resolution at the level of the four German TSOs, the data for 2025 is obtained from the SMARD platform, operated by the German Federal Network Agency (Bundesnetzagentur) [12]. This open-access dataset provides time series of electricity consumption and residual load for the four German TSOs. The original data is available as energy with a 15-minute resolution and is aggregated to hourly values.

In addition, a second dataset with spatial granularity at the level of the German federal states is employed. The data for the year 2024 is obtained from the Fraunhofer Institute for Solar Energy Systems [25]. This dataset contains time series of the average generation power per federal state and per generation technology. To derive the residual load time series at the federal state level, the generation data is combined with consumption data from the SMARD platform. Since consumption data is not directly available at the federal state resolution, auxiliary information is used to approximate the allocation. This includes total annual electricity consumption per federal state, federal state surface areas, and the geographical overlap between federal states and TSO zones [26], [27]. Based on these factors, electricity demand is redistributed proportionally to estimate state-level consumption. The residual load time series for each federal state is then calculated as the difference between estimated consumption and recorded generation.

B. Data Center Traffic Load Data

In addition to energy data, estimates of computational demand across different data centers are required for our assessment. To approximate this demand, we use traffic measurements from the DE-CIX Frankfurt Internet Exchange [28] for a representative weekday (July 08, 2025) exhibiting typical weekday load characteristics. The traffic data are aggregated to hourly averages, resulting in a 24-hour load profile. This daily profile is repeated over the analysis period of one year, as overall network traffic is generally only weakly affected by seasonal variations. The objective is to obtain a representative daily workload pattern while minor inaccuracies do not impact the general statement achieved in our work.

The observed traffic volume is converted into computational load using two parameters: the average data volume per job

and the average job duration. This enables the representation of traffic in terms of discrete computational tasks and allows the evaluation of different use cases by adjusting these parameters, for example larger job sizes for machine learning workloads. While this method simplifies network traffic to a homogeneous stream of jobs, actual traffic is highly heterogeneous. To mitigate the impact of this abstraction, the modeled system capacity is scaled according to average data center sizes reported in the literature [3].

IV. Methodology

This chapter presents a model of a system of distributed data centers, for our assessment of possible power flexibilities and CO_2 savings. We introduce an optimization problem that optimizes workload allocation between these distributed data centers with regards to residual load and quantifies the resulting emission reductions and latency trade-offs.

A. System Model

We first model per-server power consumption, then describe the system scaling and the quantification of additional delay.

1) Server Power Consumption: Server energy models typically relate power consumption to CPU utilization. Studies demonstrate that idle servers often consume 40–60 % of their peak power, with an approximately linear increase per additional active core [29]. Therefore, we assume a general server model: each server consists of 32 cores, able to process up to 32 jobs in parallel. Idling, the server consumes 50 % of its full load power draw. From there, power draw increases linearly with each active core, up to the maximum power draw of 400 W. The parameters are chosen as proxies and can be substituted for the evaluation of real world data centers. We choose fixed values instead of including them in a parameter study, since we assume all servers and data centers to be uniform, and analyze system scaling with regards to the traffic as the main influencing parameter.

Cooling and supporting infrastructure is based on the power usage effectiveness (PUE). We assume a PUE of 1.46, as reported in literature for the average German data center [30]. An extension to a more refined cooling model is not the main focus of this work and can be tackled in future work.

In our power model, we focus on server-side energy consumption and do not explicitly account for the energy consumed by network infrastructure. We exclude network energy costs because the network is assumed to be always provisioned, shared across many services, and largely dominated by idle power draw, making it difficult to attribute marginal energy consumption to individual workload placements or traffic shifts. Future work includes augmenting this model by associating data transfers with additional energy consumption.

2) System Scaling: Absolute power consumption of the system depends highly on the system scaling, which is affected by multiple parameters, e.g., number of cores per server, total traffic, and job size and duration. Instead we will focus in our evaluation on the data center capacity relative to the peak traffic volume. Depending on the use case of the data center, different scaling approaches exist. Google reports a very high server utilization of up to 80 % for web indexing, as higher utilization increases efficiency [31]. The DE-CIX Frankfurt Internet Exchange reports scaling according to peak traffic load, expanding as soon as 63 % is reached [32]. We adapt this approach in our analysis. Node capacity is treated as a parameter in the evaluation, and a baseline value and its derivation is outlined in the following. The dataset reports a peak traffic rate of 16.6 tbit/s. Because the data originates from the DE-CIX Frankfurt Internet Exchange, which is an exceptionally large deployment, we downscale this value by a factor of four, resulting in 4.15 tbit/s. This scaling does not affect the generality of the results, as the analysis is performed relative to peak load. In addition, we discretize the workload using a job duration of 100 ms and an average volume of 160 kB per job, reflecting short service requests typical for interactive applications. To achieve a peak utilization of 50 %, the system needs a total capacity of

$$\frac{2 \cdot 4.15 \frac{\text{tbit}}{\text{s}} \cdot 3600 \frac{\text{s}}{\text{h}}}{1.28 \cdot 10^{-6} \frac{\text{tbit}}{\text{job}}} = 23,343.75 \cdot 10^6 \frac{\text{jobs}}{\text{h}}.$$

Each server can process 32 jobs in parallel. With an assumed processing time of 100 ms per job and uniformly distributed arrivals over the hour, the per-server throughput is

$$32 \cdot \frac{3600 \frac{\text{s}}{\text{h}}}{0.1 \frac{\text{s}}{\text{jobs}}} = 1.152 \cdot 10^6 \frac{\text{jobs}}{\text{h}}.$$

Consequently, the system requires

$$\left\lceil \frac{23,343.75 \cdot 10^6 \frac{\text{jobs}}{\text{h}}}{1.152 \cdot 10^6 \frac{\text{jobs}}{\text{h}}} \right\rceil = 20,264$$

servers, distributed across all nodes. As introduced later, we investigate scenarios with two, four, and 16 nodes, meaning 10132, 5066, and 1266 servers per node. To contextualize this relative to typical German data centers, we derive the average number of servers based on figures reported by the BMWK report [3], which states that Germany has more than 2,000 data centers with a combined installed capacity exceeding 2,700 MW. This leads to 1.35 MW per data center. Assuming a PUE of 1.46 and a maximum power draw of 400 W per server, and not considering additional IT infrastructure, this means, the average German data center has

$$\frac{1.35\,\text{MW}}{1.46} : 400\,\text{W} \approx 2312$$

servers. Consequently, our scenarios vary from large to small scale data centers.

3) Additional Delay: Load shifting introduces additional communication latency, comprising processing delay, transmission delay, and propagation delay. Processing delay occurs at intermediate nodes, transmission delay when data is placed on links, and propagation delay as the signal travels through the medium. Additional to the increased latency for the moved data itself, moving large amounts of traffic can lead to congestion on network links and nodes, further influencing delay. To capture these effects, a detailed model

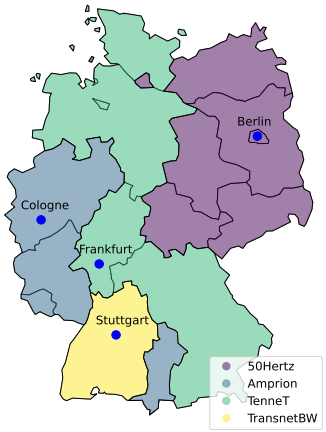

Figure 2: Four nodes scenario

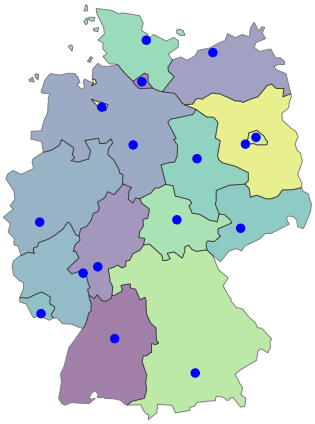

Figure 3: 16 nodes scenario

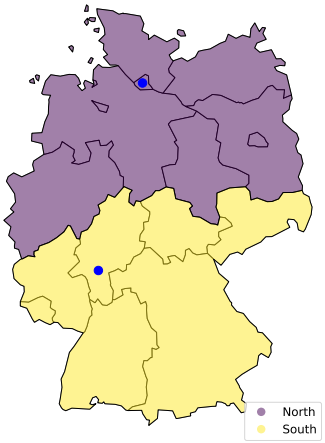

Figure 4: Two nodes scenario

of the network infrastructure is necessary. Since our objective is to quantify the additional delay introduced by geographical load shifting, we approximate end-to-end latency solely as a function of physical distance between nodes. Specifically, we model propagation delay, which scales linearly with distance and the signal propagation speed in the transmission medium. This allows us to avoid additional assumptions about network topology, queuing behavior, or hardware-specific processing characteristics, which are out of scope for this work.

Based on this abstraction, we compute a matrix of inter-regional delays based on the great-circle distance $dist_{i,j}$ in kilometers between node locations i and j. The signal propagation time through optical fiber is calculated assuming a transmission speed of 200,000 km/s. The total delay $d_{i,j}$ in milliseconds between node i and j is thus expressed as

$$d_{i,j} = \frac{dist_{i,j}}{200 \frac{km}{ms}}. \tag{1}$$

B. Formulation as Optimization Problem

The optimization model determines the hourly spatial allocation of computational load among nodes in different regions to achieve two goals: (1) maximize utilization of renewable energy surpluses, and (2) minimize the total additional delay.

The optimization is performed for each hour. Let $m_{i,j}$ represent the number of computational jobs moved from node i to node j. For numerical stability, we convert to the unit millions of jobs and implement $m_{i,j}$ as a float. Each node starts with an initial workload L_i and a capacity C_i. The capacity is constant for all hours, the workload changes. The following constraints are applied:

$$\sum_j m_{i,j} \leq L_i, \qquad \forall i \tag{2}$$

$$\sum_i m_{i,j} \leq C_j - L_j + \sum_i m_{j,i}. \qquad \forall j \tag{3}$$

The first constraint ensures that no region can offload more jobs than it has available, while the second ensures that destination nodes do not exceed their capacity.

A two-stage optimization process is applied. In the first stage, the model aims to mitigate grid congestion by moving as much data center load as possible towards nodes in zones with negative residual load, denoted as R_j for each node j, with the optimization goal

$$\max \sum_{i,j:R_j<0} m_{i,j}. \tag{4}$$

In the second stage, the solution is optimized to maintain the same total amount of relocated workload, while minimizing the cumulative network delay

$$\min \sum_{i,j} m_{i,j} \cdot d_{i,j}. \tag{5}$$

This two-stage approach ensures that grid supportive and environmental benefits are fully exploited while minimizing performance degradation. Results from this approach are compared to a weighted goal optimization and no differences are found, as both result in the optimal solution. The optimization problem is implemented using the *Pyomo* modeling environment in Python and solved using the *GLPK* linear programming solver.

C. Scenarios

We evaluate three spatial configurations of the data center network. The first configuration, illustrated in Figure 2, consists of four nodes, each located within one of the TSO zones indicated by the respective colors. The nodes are placed in Berlin, Cologne, Frankfurt, and Stuttgart. These cities represent the largest urban centers within their respective zones, with the exception of Frankfurt, which is selected specifically due to its high concentration of data centers in practice. Note that TSO zones are only roughly drawn in the figure, as the real borders do not follow federal state borders. This scenario is particularly relevant, as it best reflects realistic data center placement. Operators typically select locations near core network nodes, which under typical topology assumptions are placed as four to five nodes in large German cities [33]. The second configuration, shown in Figure 3, refines spatial

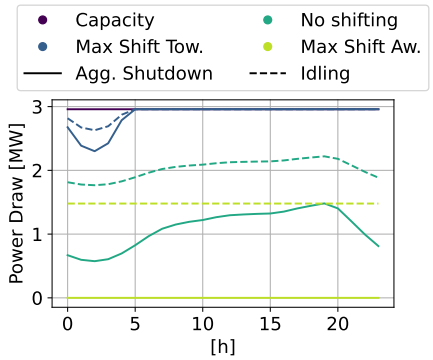

Figure 5: Flexibilities in power draw of one node in four nodes scenario.

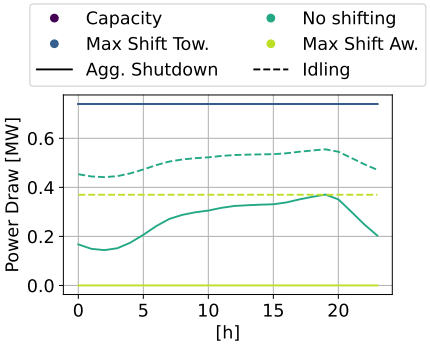

Figure 6: Flexibilities in power draw of one node in 16 nodes scenario.

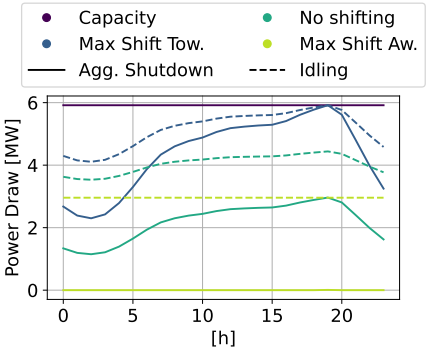

Figure 7: Flexibilities in power draw of one node in two nodes scenario.

Table I: CO_2 factors for electricity generation [34], [35]

Generation	kgCO_2e/MWh	Generation	kgCO_2e/MWh
Biomass	230	Lignite	820
Hydropower	24	Hard Coal	820
Wind Offshore	12	Natural Gas	490
Wind Onshore	11	Oil	300
Photovoltaics	48	Geothermal	38
Pump Storage	24	Waste	300
Other Renewables	24	Other Conventionals	710

resolution to 16 nodes, assigning one node to each German federal state, allowing for a more fine-granular evaluation of especially CO_2 savings. This scenario is chosen, since one of the four TSO's covers a large area, spanning from the north of Germany to the south, possibly obscuring congestion effects. The third scenario aggregates the federal states into two macro-zones representing northern and southern Germany, chosen to illustrate the north-south congestion.

For each spatial scenario, we analyze two operational modes: (i) a baseline mode in which all servers remain idling when unoccupied; (ii) an aggressive server-shutdown mode in which all unused servers are powered down, including their cooling. These scenarios represent lower and upper bounds on the achievable flexibility and energy savings. Realistically, maintaining a buffer of idling servers for resilience and rapid response is necessary, meaning results lie within these bounds.

D. Target Metrics

We evaluate our optimization problem using two target metrics, representing the potential flexibilities for the energy grid, and the tradeoff between sustainability and service quality.

1) Flexibilities in Power Draw: To characterize the power flexibility offered by a data center, we visualize how much power can be shifted to other data centers during a typical day. These flexibilities for a data center location are limited by the available load and the excess capacity in other locations, and investigate the potential of including data centers in mitigation of energy grid congestion.

2) CO_2 Emissions: To translate power consumption of the data centers into CO_2 emissions, we first calculate CO_2 intensity of electricity generation for our dataset using the weighted average

$$I_{z,t} = \frac{\sum_k G_{z,t,k} \cdot f_k}{\sum_k G_{z,t,k}}, \tag{6}$$

where $G_{z,t,k}$ denotes the generation in megawatt-hours from technology k in zone z for timeslot t with f_k as its emission factor (kgCO_2e/MWh). Emission factors are taken from [34] and [35] and are shown in Table I. We note that this approach relies on zonal average generation and does not account for cross-border electricity exchanges. Thus, the resulting CO_2 emissions should be interpreted as an approximation based on local generation mixes, which is justified in case of energy grid congestion where shifting load closer to renewable generation helps alleviate grid stress.

V. Evaluation

This section evaluates the grid-supporting flexibility and CO_2 savings of geographically distributed data centers.

A. Flexibilities in Power Draw

We start by modeling the flexibilities in the power draw of a single data center, which can be utilized and monetarized for redispatch purposes.

Figure 5 illustrates power draw flexibility for a single node in the four-node scenario. The x-axis shows hours of the day, and the y-axis shows power in megawatts. Purple indicates power draw while operating at maximum capacity, teal shows power draw fluctuating with actual load, blue shows power if all traffic from other nodes is shifted in, and green shows power if the node offloads all its traffic. Solid lines represent aggressive shutdown mode with idle servers off, while dashed lines show idling mode with idle servers running. These constitute a lower and upper bound on the actual power draw. Figure 6 and Figure 7 present the same analysis for the 16-node scenario and the north-south scenario, respectively. In all plots, potential flexibility can be read as the differences between the lines. The highest flexibility can be achieved between a maximum shift towards a node and the maximum shift away, essentially turning this node off completely. Depending on the time of the day, the concrete

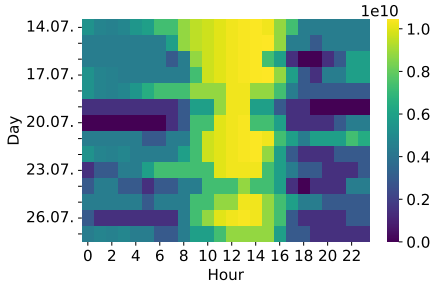

Figure 8: Total moved traffic during timeslots for two weeks in summer in 16 nodes scenario.

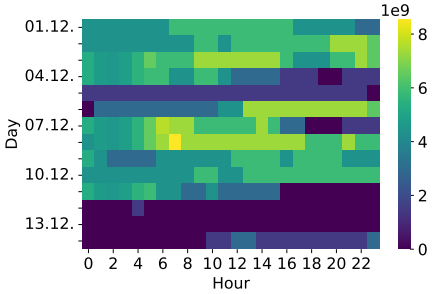

Figure 9: Total moved traffic during timeslots for two weeks in winter in 16 nodes scenario.

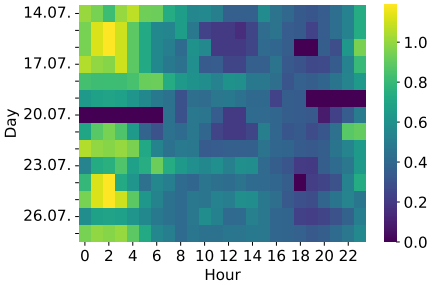

Figure 10: Mean propagation delay [ms] for moved traffic during timeslots for a week in summer in 16 nodes scenario.

operation mode, and the scaling of the data centers, this value varies. In our analysis we achieve a maximum flexibility of 6 MW for the two nodes scenario at peak traffic hour between the maximum shift towards and maximum shift away in the aggressive shutdown operational mode. Considering the planned integration of smaller flexibility resources into redispatch processes and the ongoing decentralization of the energy system, this magnitude can provide a meaningful contribution to redispatch operations.

The results demonstrate that the aggressive shutdown scenario results in lower power consumption than the idling mode. When servers are idling, the system has an unflexible base electric load. Shutting off servers reduces energy consumption, but has a trade-off with service availability and system resilience. The flexibility is influenced by the capacity of each node and the system workload. Higher node capacity allows more shifted traffic, increasing flexibility. This effect is evident in both the four-node and 16-node scenarios, where the blue "maximum shift" lines are limited by node capacity during daytime hours. In the north-south scenario system traffic rather than node capacity is the limiting factor, as the blue lines reach the maximum capacity only during peak hour. Workload patterns also play a critical role. This analysis assumes a typical day-night cycle, but flexibility may differ under alternative usage patterns. Additionally, while it is possible to artificially increase flexibility by keeping servers busy with unnecessary tasks, doing so results in wasted energy.

With this we can answer our first research question, RQ1: *Data centers can provide flexibilities for the German energy grid in meaningful magnitudes, in our scaling and example modeling up to 6 MW. However, the concrete value depends on the theoretical assumptions in this work and should be recalculated with input from real data centers. The flexibilities depend on the time of the day, the operational mode, and the concrete available infrastructure. They are limited by the available workload and the excess capacity.*

B. Optimization Impact on Traffic

To investigate the effect of the optimization, Figure 8 and Figure 9 visualize the amount of moved traffic. Both figures show the 16-node scenario across two weeks: Figure 8 in summer and Figure 9 in winter. Both figures reflect typical renewable generation and electrical demand patterns. In summer, traffic shifts mainly around noon, when solar output is high and both work load and grid load are moderate, while little shifting occurs in the evening due to declining solar generation and rising household power demand. Optimization is feasible on all days. In winter, shifting occurs primarily at night, driven by wind generation. However, from December 11 onward, no optimization is possible for several days, likely due to a weather condition known as dunkelflaute, with no wind and solar production.

Figure 10 shows the mean additional propagation delay for shifted traffic in the 16-node scenario over two summer weeks. As delay is modeled distance-based, it reflects the extent of geographic shifting. Delays are higher at night due to wind-driven shifting, while noon traffic, though larger in volume, incurs moderate delay. Note that this abstraction excludes processing and transmission delays and serves only to indicate shifting distances. Detailed delay and congestion effects require further study.

Consequently, we answer RQ2 with: *Workload shifting is most effective during periods of high renewable energy availability and moderate energy demand: Around midday in summer, when solar production peaks, and during nighttime in winter, when wind energy dominates. Low utilization periods further increase shifting potential, while weather conditions without wind or sun limit optimization opportunities regardless of time. During wind dominated production, traffic needs to be shifted further in Germany, increasing the additional delay.*

C. Influence of System Scaling and Potential CO_2 Emissions

After optimizing workload placement with renewable production, we assess relative potential CO_2 emission savings. Since spare capacity limits optimization, we first examine how system scaling impacts the amount of shifted workload.

Figure 11 shows the total number of moved jobs versus peak utilization, reflecting system dimensioning relative to peak demand. Lower peak utilization indicates greater overprovisioning and excess capacity. The four node scenario is shown in blue, the two node scenario in purple, and the 16 node scenario in green. Overall, lower peak utilization enables more workload shifting, most noticeably in the 16 node scenario, where the higher geographic resolution creates more

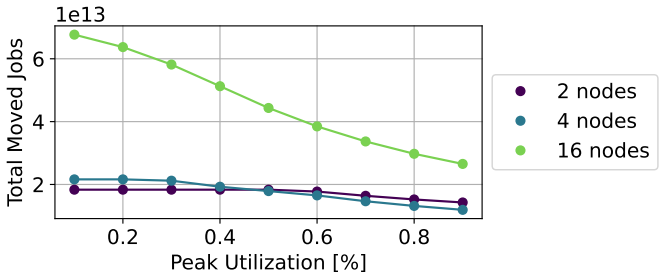

Figure 11: Effect of system scaling on amount of moved traffic for each geographic scenario.

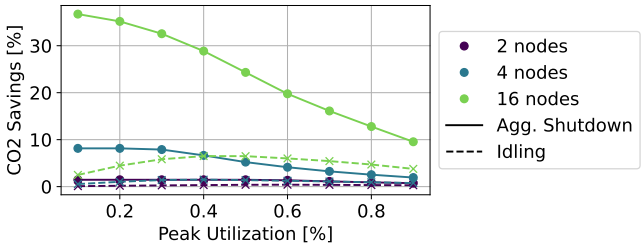

Figure 12: Effect of system scaling on relative CO_2 savings for each scenario combination.

time windows with negative residual load and therefore more optimization opportunities. In contrast, the two node and four node scenarios reach a saturation point, where additional over-provisioning yields little further increase in shifted workload, at approximately 30 % and 60 % peak utilization, respectively.

Building on this, we evaluate the achievable relative CO_2 savings. Figure 12 shows relative CO_2 reduction versus peak utilization for all geographic scenarios and operational strategies. Peak utilization is plotted on the x-axis and relative CO_2 savings on the y-axis. For each scenario, the two operational strategies are evaluated: the idling strategy, in which unused servers remain powered but idle (dashed lines), and the aggressive shutdown strategy, in which idle servers are turned off immediately (solid lines). Note that the server shutdown strategies apply before and after the optimization, i.e., CO_2 emissions are compared to the before optimization state, where unneeded servers are also shut down. Therefore, saving potential indeed comes from workload shifting and turning off servers at a less beneficial location, moving the load towards better locations, not just turning off more servers.

In case of aggressive shutdown, the 16 node scenario consistently achieves the highest reductions of up to 37 % in strongly over-dimensioned systems and about 10 % for a peak utilization of 90 %. The four node scenario shows moderate savings below roughly 10 %, while the two node scenario yields only limited reductions. As peak utilization decreases, more workload can be shifted away from locations with high residual load and carbon intensity, and more servers can be powered down entirely in these regions.

It is important to interpret these results carefully. We only account for operational emissions. Increasing installed capacity implies additional embodied CO_2 emissions from hardware manufacturing, construction, and deployment, which are not considered here. Very low peak utilization also implies inefficient operation before the optimization, which already leads to a higher saving potential. Moreover, negative residual load does not necessarily imply renewable curtailment or redispatch, as sufficient transmission capacity may still be available. We therefore use residual load as a practical indicator due to data availability, while a precise quantification of the actual CO_2 impact would require actual redispatch data and power flow modeling, which is beyond the scope of this study. Nevertheless, the results show that meaningful operational CO_2 savings are achievable not only in heavily over-dimensioned systems but also at comparatively high peak utilization, with less excess capacity.

For the idling strategy, the relationship differs. Here, maximum relative savings occur at intermediate peak utilization (around 50 %). At very low peak utilization, meaning the same traffic is handled by more servers, a large share of total power draw is dominated by idle consumption from excess servers, which cannot be eliminated without shutdown. As a result, workload shifting alone yields limited relative benefit. As peak utilization rises beyond 50 %, both the number of excess servers and the total amount of shifted workload decrease, reducing the achievable savings again.

Finally, this answers RQ3: *Distributed data centers can reduce operational CO_2 emissions through workload shifting, with savings depending on geographic granularity and spare capacity. Using residual load as a proxy for excess renewable generation, we observe reductions of up to 37 % in highly distributed, over-provisioned deployments with aggressive server shutdown, and around 10 % even at 90 % peak utilization. Savings are lower in coarse-grained deployments and when servers remain idle. However, for a concise analysis of CO_2 savings, additional grid data is necessary.*

VI. Discussion and Conclusion

The increasing decentralization of renewable electricity generation poses growing challenges for grid stability and congestion. Distributed data centers, due to their spatial workload flexibility, represent a promising demand-side resource to better align electricity consumption with renewable availability.

In this work, we presented a data-driven model linking geographically distributed data centers with renewable generation and regional CO_2 intensity in Germany. The results depend on theoretical assumptions, and aim to show general effects rather than concrete numbers. They show that distributed data centers can provide meaningful grid-side flexibility, reaching several megawatts per location depending on infrastructure and operating mode. Shifting benefits are time-dependent and peak during high renewable availability and moderate demand, notably midday in summer and nighttime in winter. Additionally, renewable-aware workload shifting may enable significant CO_2 reductions of up to 37 % in highly distributed, over-provisioned scenarios with aggressive server shutdown,

and about 10 % even at high utilization. However, our study uses assumptions and residual load as a proxy for renewable surplus, while more detailed modeling of transmission bottlenecks is needed to estimate real-world CO_2 reduction potential.

The practical potential of workload shifting depends on workload characteristics and operational strategy. Time-varying, underutilized workloads offer the greatest flexibility and CO_2 savings, while persistently high utilization limits shifting potential. Operational choices further matter: keeping idle servers running preserves availability but increases consumption, whereas shutting them down in high-CO_2 regions reduces emissions. Thus, flexibility and environmental benefits depend on both workload and idle resource management.

Implementing load shifting requires coordination between the energy and communication sectors. Accurate forecasts of network traffic and renewable generation must be exchanged, e.g., via regional electricity prices, enabling data centers to make cost-driven shifting decisions. However, practical challenges remain: latency constraints, data locality, and service-level agreements limit migration. Frequent relocation introduces orchestration overhead and may increase network congestion, and server power cycling can reduce hardware lifespan and react too slowly to short-term grid dynamics.

Future work will focus on incorporating realistic data center workloads, extending the analysis to additional countries, and refining the system model with improved network modeling to better capture delay and congestion.

Acknowledgments

This work was partly funded by the German Federal Ministry of Research, Technology and Space Grant 16KIS2282 of the University of Würzburg and Grant 16KIS2281 of the University of Stuttgart ("SUSTAINET-Advance").

References

[1] I. Rhoden *et al.*, "Spatial Heterogeneity-Challenge and Opportunity for Net-Zero Germany," 2021.

[2] W.-P. Schill, "Residual Load, Renewable Surplus Generation and Storage Requirements in Germany," *Energy Policy*, 2014.

[3] R. Hintemann *et al.*, "Status and Development of the German Data Centre Landscape – Executive Summary." BMWK, 2024.

[4] Umweltbundesamt. (2025) Stromverbrauch. Accessed: 2026-02-18. [Online]. Available: https://www.umweltbundesamt.de/daten/energie/stromverbrauch

[5] T. Hoßfeld, "Energy Use in Data Centers: Current Figures and Trends," 2025.

[6] M. Schwenke *et al.*, "Review of Concepts for Operational Congestion Management with Regard to Redispatch 2.0," in *CIRED 2024 Vienna Workshop*.

[7] F. B. Marten *et al.*, "Redispatch 3.0-Optimierung von Kleinstflexibilitäten," in *Fachtagung Hochautomatisierter Netzbetrieb 2024*, 2024.

[8] D. Bauknecht *et al.*, "The Role of Decentralised Flexibility Options for Managing Transmission Grid Congestions in Germany," *The Electricity Journal*, 2024.

[9] F. Loh *et al.*, "Qos and Qoe Study of the European 5G Mobile Networks for Next Generation of Applications," *Communications Magazine*, 2025.

[10] O. O. Yolcan, "World Energy Outlook and State of Renewable Energy: 10-Year Evaluation," *Innovation and Green Development*, 2023.

[11] Bundesnetzagentur, "2024 Electricity Market Data," accessed: 2025-12-03. [Online]. Available: https://www.bundesnetzagentur.de/SharedDocs/Pressemitteilungen/EN/2025/20250103_SMARD.html

[12] "SMARD – Strom- und Gasmarktdaten," https://www.smard.de/home, Bundesnetzagentur, 2026, accessed: 2026-02-05.

[13] International Energy Agency (IEA), "Grid Congestion is Posing Challenges for Energy Security and Transitions – Analysis," 2025. [Online]. Available: https://www.iea.org/commentaries/grid-congestion-is-posing-challenges-for-energy-security-and-transitions

[14] Energy Systems of the Future (ESYS), "Grid Congestion as a Challenge for the Electricity System: Options for a Future Market Design," German National Academy of Sciences Leopoldina and acatech - National Academy of Science and Engineering and Union of the German Academies of Sciences and Humanities, Tech. Rep., 2021. [Online]. Available: https://www.akademienunion.de/fileadmin/au-uploads/publikationen/Publikationen_PDFs/2021/2021_Position_Paper_ESYS_Grid_Congestion.pdf

[15] A. Rahman, X. Liu, and F. Kong, "A Survey on Geographic Load Balancing Based Data Center Power Management in the Smart Grid Environment," *IEEE Communications Surveys & Tutorials*, 2013.

[16] J. Schiller and M. Pruckner, "Sustainable, Situation-Aware Multi Objective Spatial Load Scheduling for Data Centers," in *16th ACM International Conference on Future and Sustainable Energy Systems*, 2025.

[17] H. Hu *et al.*, "Coordinating Workload Scheduling of Geo-Distributed Data Centers and Electricity Generation of Smart Grid," *IEEE Transactions on Services Computing*, 2017.

[18] C. Li *et al.*, "Managing Server Clusters on Renewable Energy Mix," *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*, 2016.

[19] Z. Liu *et al.*, "Greening Geographical Load Balancing," *IEEE/ACM Transactions on Networking*, 2014.

[20] B. Camus *et al.*, "Harnessing the Geographical Flexibility of Distributed Computing Clouds for Cooperative Self-Consumption," in *2018 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe)*. IEEE, 2018.

[21] A. Agarwal *et al.*, "Redesigning Data Centers for Renewable Energy," in *20th ACM Workshop on Hot Topics in Networks*, 2021.

[22] M. Diamanti *et al.*, "Incentive Mechanism and Resource Allocation for Edge-Fog Networks Driven by Multi-Dimensional Contract and Game Theories," *IEEE Open Journal of the Communications Society*, 2022.

[23] K. Nguyen *et al.*, "FaaSt: A Latency-Aware Serverless Scheme for Edge-Cloud Environments," in *2024 IEEE 13th International Conference on Cloud Networking*, 2024.

[24] ——, "Investigation of Serverless Consumption and Performance in Multi-Access Edge Computing," in *2024 International Conference on Information Networking (ICOIN)*. IEEE, 2024.

[25] Fraunhofer Institute for Solar Energy Systems. (2026) Accessed: 2026-01-30. [Online]. Available: https://www.energy-charts.info

[26] Statistisches Bundesamt (Destatis). (2026) Regionaldatenbank Deutschland (GENESIS-Online) - Eintrag „86231". Accessed: 2026-02-18. [Online]. Available: https://www.regionalstatistik.de

[27] Statistische Ämter des Bundes und der Länder. (2026) Fläche und Bevölkerung. Accessed: 2026-02-18. [Online]. Available: https://www.statistikportal.de/de/bevoelkerung/flaeche-und-bevoelkerung

[28] DE-CIX Management GmbH, "Frankfurt Traffic Statistics," DE-CIX Management GmbH, accessed on 2025-01-28. [Online]. Available: https://www.de-cix.net/de/standorte/frankfurt/statistiken

[29] E. Ahvar, A.-C. Orgerie, and A. Lebre, "Estimating Energy Consumption of Cloud, Fog, and Edge Computing Infrastructures," *IEEE Transactions on Sustainable Computing*, 2019.

[30] BMWK, "Stand und Entwicklung des Rechenzentrumsstandorts Deutschland Gutachten im Auftrag des Bundesministeriums für Wirtschaft und Klimaschutz," 2025.

[31] M. Kanellos. (2013) Google Says: Save Energy, Ditch Your Data Center. Forbes. Accessed: 2026-02-02. [Online]. Available: https://www.forbes.com/sites/michaelkanellos/2013/06/06/google-says-save-energy-ditch-your-data-center/

[32] DE-CIX Management GmbH, "Internet Exchange Operator DE-CIX Sees a Strong Change in Internet User Behavior," https://www.de-cix.net/en/about-de-cix/media/press-releases/internet-exchange-operator-de-cix-sees-a-strong-change-in-internet-user-behavior, 2020, accessed: 2026-02-02.

[33] F. Poignée *et al.*, "POBTOG: A Population-Based Topology Generator for Country-Wide Communication Networks," *NetSys*, 2025.

[34] S. Schlömer *et al.*, "Annex III: Technology-Specific Cost and Performance Parameters," in *Climate Change 2014: Mitigation of Climate Change: Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, 2014.

[35] NREL, "Life Cycle Greenhouse Gas Emissions from Electricity Generation: Update."