energies

MDPI

Review

Large-Scale Data Centers in High-Renewable-Penetration Power Systems: A Review of Flexible Resources, Grid Interactions, and Operational Impacts

Maosen Chen [1], Wenzhen Li [2,*], Shuai Mao [3] and Hui Ma [4]

[1] School of Electronic and Information, Xi'an Polytechnic University, Xi'an 710600, China; 42303040301@stu.xpu.edu.cn
[2] School of Aeronautical Manufacturing and Vehicle Engineering, Xihang University, Xi'an 710077, China
[3] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China; maoshuai@nwpu.edu.cn
[4] Aviation Key Laboratory of Science and Technology on Aero Electromechanical System Integration, Nanjing 211106, China; valen@nuaa.edu.cn
* Correspondence: 202212004@xaau.edu.cn

Abstract

This paper reviews large-scale data centers in high-renewable-penetration power systems from the perspectives of flexible resources, grid interactions, and operational impacts. First, this paper summarizes the electrical and operational characteristics of large-scale data centers and renewable-rich power systems, and identifies the main coupling challenges between them. Second, this paper reviews the major flexible resources available in data centers, including workload flexibility, cooling and thermal flexibility, uninterruptible power supply systems, battery energy storage systems, and backup power resources. Third, this paper examines the main interaction pathways between data centers and power systems, with emphasis on demand response, ancillary services, and multi-timescale coordination. Finally, this paper discusses the operational impacts of these interactions on grid operation, reliability, service quality, economic performance, and low-carbon operation. The review could provide a structured understanding of how large-scale data centers can evolve from passive loads into flexible and grid-supportive resources under renewable-rich grid conditions.

Keywords: large-scale data centers; high-renewable-penetration power systems; flexibility resources; grid interactions; operational impacts

Check for updates

Academic Editor: Alan Brent

Received: 23 June 2026
Revised: 14 July 2026
Accepted: 22 July 2026
Published: 28 July 2026

1. Introduction

In recent years, power systems with a high share of renewable energy have become a major trend in modern energy systems. These renewable sources mainly include wind power, solar power, hydropower, and, in some areas, other low-carbon resources connected through power converters [1–3]. Power systems usually connect these resources through large wind and solar plants, distributed generation, microgrids, and different kinds of power electronic devices. A higher share of renewable energy can improve sustainability and help reduce carbon emissions. At the same time, it also changes how the power grid operates. The grid faces stronger power fluctuations [4]. The system inertia becomes lower. The need for flexibility also becomes more obvious, especially in power balancing, frequency control, and reserve support [5]. Meanwhile, large-scale data centers are becoming one of the fastest-growing types of electricity load, as shown in Figure 1. These data centers usually have high power density, fast expansion speed, and very strict demands on power supply reliability and power quality [6,7]. When people place data centers in power systems with abundant renewable energy, these facilities can also bring some clear benefits. For example, they can support flexible load control, help coordinate on-site energy storage and backup resources, and improve renewable energy use through smart scheduling. However, the close connection between large-scale data centers and high-renewable power systems also brings new problems [8]. This coupling can make net-load fluctuations more severe, which can increase stress on local networks. It can also make coordination across different time scales more difficult. In addition, there is growing conflict between the flexibility that the grid needs and the strict service requirements that data centers must meet [9].

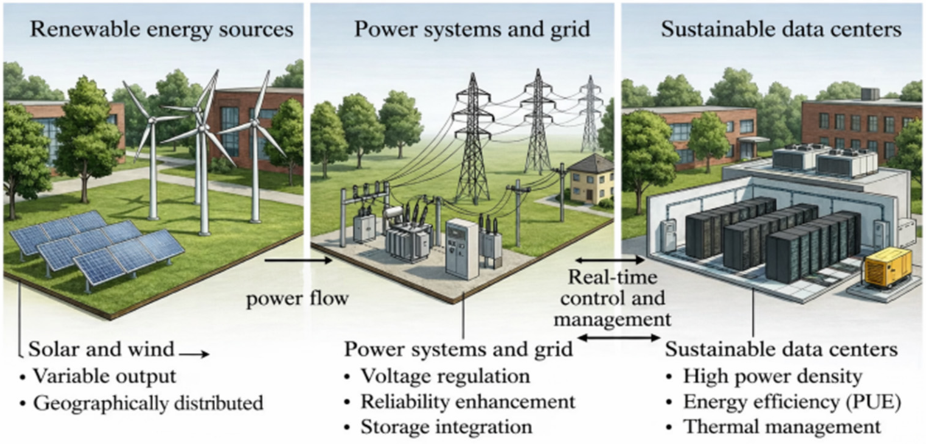

Figure 1. Interactions of large-scale data centers and high-renewable-penetration power systems.

Recent studies have gradually pointed out that when large-scale data centers are deeply connected with power systems that have a high share of renewable energy, this connection is not only a chance to reduce carbon emissions. It also brings new problems in system operation. One key issue comes from the fast growth of data-center load [8]. This problem is more obvious in AI computing scenarios. The rising load can make net-load changes more severe. It can also increase the pressure on local transmission and distribution networks. This problem becomes even more serious when data centers are connected to feeders that are already heavily loaded or to local grids with limited support capacity. Another important topic is the difficulty of coordination over different time scales. To deal with this issue, existing studies have started to build multi-time-scale scheduling and joint optimization methods [10,11]. These methods usually consider renewable power changes, workload transfer, energy storage control, and grid service tasks at the same time. The results show that time flexibility and location flexibility should work together. Researchers should not study them separately. A third problem is the increasing tension between grid support and data-center service needs. Ref. [12] shows that data centers can help the grid through demand response, storage support, and auxiliary services. However, this kind of flexibility is always limited by service agreements, service quality needs, delay sensitivity, and reliability requirements [13]. Overall, large-scale data centers have strong potential to help power systems with high renewable energy use. They can do this through flexible operation and the coordinated use of different energy resources. However, this potential cannot be fully used without careful coordination. Researchers need to balance the flexibility needs of the power grid with the strict requirements of data center operation, especially in performance and service quality.

In recent years, a small but fast-growing number of review papers have started to examine the link between data centers and power systems. At least five recent review studies can be found in related areas. However, their focus is still quite scattered. For example, ref. [8] mainly studies data centers as providers of flexibility. It discusses demand response, adjustable workloads, and possible ways for data centers to join electricity markets and support power systems with high shares of renewable energy. In [10], energy infrastructure and storage technologies for grid connection are considered, which focuses on uninterruptible power supply (UPS) systems, battery energy storage systems, grid-interactive UPS, and hybrid storage designs for AI-driven data centers. In [14], sustainable data center operation and integrated smart energy management are discussed, which covers renewable energy supply structures, power delivery systems, thermal management, and low-carbon operating methods. In addition, references [15,16] discuss AI-oriented data centers as power users that can also interact with smart grids and district energy systems. They highlight the special workload features of such data centers and their possible flexibility value. Although these studies offer useful insights, they still lack a unified discussion of large-scale data centers in power systems with high renewable penetration. In particular, most existing work does not examine this topic together with the three linked aspects of flexible resources, grid interaction, and operational impact. More specifically, current reviews give limited discussion to several key issues. One issue is how different flexible resources inside data centers work with renewable-rich grids across multiple time scales. Another issue is how these large facilities influence local grid operation and system stability. A further issue is how grid-support actions can be coordinated with strict requirements on service level, reliability, and computing performance. For this reason, a dedicated review is still needed. Such a review should connect these three aspects in a clear and systematic way.

As shown in Table 1, previous reviews mainly focus on individual aspects, such as workload flexibility, energy storage, sustainable operation, AI-oriented facilities, or district-energy integration. In contrast, the present review connects internal flexibility resources, external grid-interaction mechanisms, and operational impacts within a unified power-system-oriented framework. This integrated resource–interaction–impact perspective constitutes the main contribution of this review.

Table 1. Comparison of previous reviews and the scope of the present paper.

Review	Main Focus	Flexibility Resources	Grid Interaction	Operational Impacts
Takci et al. [8]	Data centers as flexible loads	Mainly workload flexibility	Demand response and grid flexibility	Mainly grid-side impacts
Mohammadi et al. [10]	Energy storage for AI data centers	UPS, BESS, hybrid storage	Storage-based grid support	Mainly storage and grid integration
Kumar et al. [14]	Sustainable data-center operation	Renewable energy, cooling, storage	Energy management and renewable integration	Mainly efficiency and sustainability
Zhang et al. [15]	AI-focused data centers and smart grids	AI workloads and energy resources	Smart-grid and district-energy interaction	Selected grid and thermal impacts
Huang et al. [16]	Data centers as prosumers	Renewable energy and waste heat	District-energy interaction	Mainly energy efficiency and heat recovery
Present review	Large-scale data centers in high-renewable power systems	Workload, cooling, UPS, BESS, and backup power	Demand response, ancillary services, reserves, and congestion management	Grid operation, reliability, QoS, economics, and carbon impacts

To address these issues, this paper gives a full review of large-scale data centers in power systems with high renewable penetration. The review mainly focuses on flexible resources, grid interaction, and operational impacts. Different from traditional review papers, this paper does not study data-center energy management, storage design, or flexibility support from only one angle. Instead, this paper builds a more complete framework from the viewpoint of the power system. This framework links internal data-center resources with external grid operation and the needs of renewable-rich power systems. Based on this framework, the review not only summarizes the existing flexibility paths inside large-scale data centers, but also explains how these resources interact with the grid over different time scales. The review also discusses how these interactions influence both power-system operation and data-center performance. During the preparation of this manuscript, the authors used ChatGPT 4.5 (OpenAI) for language polishing and figure improvements. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

The main contributions of this paper are as follows:

(1) An integrated classification and critical comparison of data-center flexibility resources is established. Unlike previous reviews that mainly discuss workload flexibility, storage technologies, thermal management, or sustainable operation separately, this paper places workload scheduling, cooling and thermal flexibility, UPS systems, BESSs, and backup generation within one power-system-oriented framework.
(2) The internal flexibility resources are systematically linked to external grid-interaction mechanisms, including demand response, frequency-related services, reserve provision, congestion management, and multi-timescale coordination.
(3) A bidirectional assessment of operational impacts and trade-offs is provided. The review jointly examines grid operation, local network stress, reliability, SLA and QoS requirements, economic performance, battery degradation, and carbon impacts, thereby distinguishing theoretical flexibility potential from technically feasible and operationally usable flexibility.

The organization of this paper is illustrated in Figure 2. Section 2 introduces the characteristics of large-scale data centers and high-renewable-penetration power systems, and discusses the main coupling challenges between them. Section 3 reviews the flexible resources available in large-scale data centers, including workload, cooling, storage, and backup-power flexibility. Section 4 summarizes the major grid interaction mechanisms of large-scale data centers, with emphasis on demand response, ancillary services, and multi-timescale coordination. Section 5 discusses the operational impacts of these interactions on both power systems and data centers, including reliability, service quality, economic performance, and low-carbon operation. Section 6 discusses the main challenges and future research directions. Section 7 concludes the review.

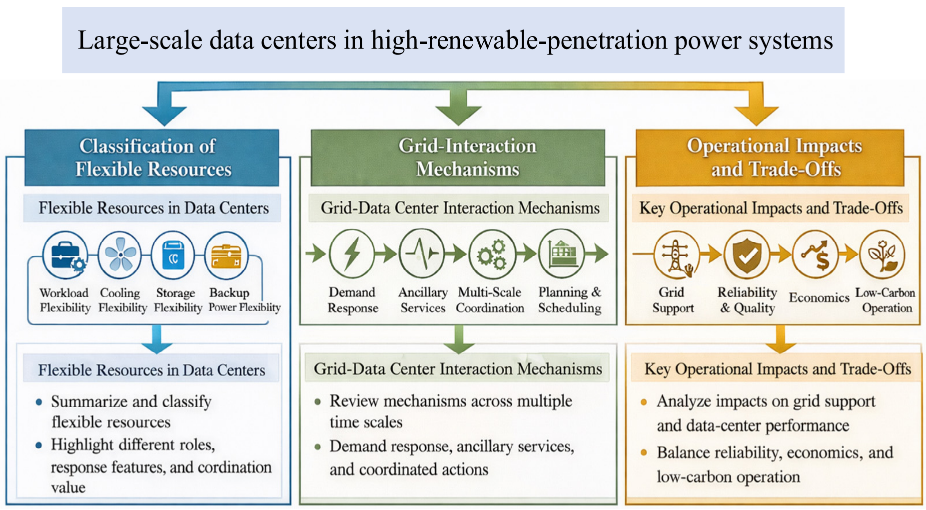

Figure 2. Review framework for data centers in high-renewable power systems.

2. Characteristics of Large-Scale Data Centers in High-Renewable-Penetration Power Systems

2.1. Electrical and Operational Characteristics of Large-Scale Data Centers

Large-scale data centers have become a special type of electricity user. This change is mainly due to their huge power demand, concentrated location, and very strict need for continuous operation, as shown in Table 2. The International Energy Agency reported that data centers used about 415 TWh of electricity worldwide in 2024. This value was about 1.5% of global electricity consumption. The same report also showed that this number may rise to about 945 TWh by 2030, and AI is expected to be the main reason for this growth [1]. At the level of a single facility, the electrical scale of new data centers has also grown quickly. North American Electric Reliability Corporation (NERC) pointed out that new large loads, including data centers, may range from several megawatts to several gigawatts. This range makes them similar to major industrial loads in terms of grid influence [17]. For this reason, large-scale data centers should no longer be viewed as ordinary commercial buildings. They should be regarded as highly concentrated electrical infrastructures whose grid connection, expansion, and operation can directly influence local and regional power systems.

Table 2. Electrical and operational characteristics of large-scale data centers.

Aspect	Main Features	Implications for Power Systems	References
Load scale and growth	• 415 TWh in 2024 • 945 TWh by 2030 • Rapid capacity expansion	• Major electricity consumer • Strong grid impact • Higher planning pressure	[1]
Large-load nature and siting	• MW-to-GW demand • Spatially concentrated loads	• Feeder loading increase • Substation stress rise • Harder connection studies	[17]
Power supply architecture	• Utility interconnection • Medium-voltage distribution • UPS systems • Backup generators • Supervisory control layers	• Higher supply continuity • Greater control complexity • Harder maintenance coordination	[6,17]
Reliability constraints	• Strict uptime targets • High power quality • Redundancy requirements • Fault-tolerant design	• Limited direct curtailment • Restricted flexibility range • Service continuity priority	[6]
Cooling–power coupling	• Cooling-related demand • IT-load dependence • Continuous heat removal	• Coupled load variation • Higher peak demand • Stronger thermal coupling	[18,19]
Increasing rack density	• Higher rack density • More thermal stress • Liquid cooling adoption	• Greater local stress • Higher cooling burden • Electro-thermal coordination	[19]
AI/HPC-driven dynamics	• AI-driven growth • Higher power density • Rapid load fluctuation • Cyclical ramping behavior • Stronger temporal variability	• Harder load forecasting • More dynamic operation • Tougher grid coordination	[1,17–19]

From the viewpoint of internal electrical structure, large-scale data centers are usually built on a layered power supply architecture. This architecture is designed to keep critical information-technology (IT) equipment running without interruption. It often includes grid connection, medium-voltage distribution, transformers, switchgear, power distribution units, UPS systems, backup generators, and supervisory control and monitoring systems [6,17]. Their operating logic is strongly influenced by reliability requirements. The Tier Classification system proposed by Uptime Institute is still one of the most widely used standards for evaluating data-center infrastructure performance. This system clearly shows that higher availability targets require higher redundancy, stronger fault tolerance, and stricter operation management. In actual engineering, this means that data-center electrical design is not decided only by efficiency and cost. It is also limited by strict requirements on outage tolerance, power quality, maintenance flexibility, and recovery capability [6].

Another key feature is the close link between electrical demand and cooling demand. In large data centers, cooling is not just a supporting subsystem. It is an essential part of the total electrical load because the heat produced by servers, network equipment, and power electronic devices must be removed continuously to keep the system operating safely. Uptime Institute reported that average rack density is rising, and high-density racks are becoming more common. American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) also pointed out that the growth of high-density computing is pushing liquid cooling into wider use in mainstream data centers [18,19]. As a result, the electrical behavior of a data center is closely tied to its thermal management method. Changes in IT load directly change cooling demand. At the same time, cooling-system design also affects total facility power, peak demand, dynamic response, and infrastructure configuration [19].

These features are becoming more obvious as AI- and high-performance-computing (HPC)-oriented data centers continue to grow. The growth of AI is increasing rack power density and thermal pressure. This trend is putting greater pressure on the power delivery and cooling ability of existing infrastructure [1,18]. At the same time, NERC warned that new large loads may show fast demand fluctuations and repeated ramping behavior. This trend creates extra difficulties for load forecasting, infrastructure planning, and real-time grid coordination. Therefore, compared with traditional data centers, AI- and HPC-oriented facilities not only consume more electricity, but they also show stronger time-varying behavior, closer electro-thermal coupling, and stricter requirements on both internal infrastructure and external grid support [17,19].

2.2. Specific Characteristics of AI-Driven Data Centers

AI-driven data centers differ from conventional enterprise and cloud data centers not only in electricity-consumption scale, but also in workload structure, power dynamics, cooling requirements, and operational flexibility. Their main computing loads are usually associated with large-scale model training and real-time inference. These two workload categories have different electrical and scheduling characteristics. Training tasks often run for long periods, involve a large number of accelerators, and maintain relatively high utilization levels. By contrast, inference workloads are more closely linked to user requests and may show faster and less predictable variations in demand. Therefore, treating all AI workloads as a single homogeneous load may lead to inaccurate estimates of both grid impact and flexibility potential.

Large AI training clusters can create high and spatially concentrated power demand because thousands of graphics processing units or other accelerators may operate simultaneously. In addition to the average power level, synchronized computing stages, checkpoint operations, and communication-intensive phases may produce rapid changes in IT demand. These variations are transmitted to the cooling and power-supply systems, thereby increasing short-term ramping, local transformer loading, and power-quality requirements. When several AI facilities are connected within the same region, correlated load growth or synchronized operational behavior may further increase pressure on substations, feeders, and regional generation adequacy [1,17].

AI-driven facilities also strengthen the coupling between computing and thermal management. High accelerator utilization produces large and highly concentrated heat loads, which encourages the use of direct-to-chip liquid cooling, immersion cooling, and other high-density thermal-management technologies. These cooling structures can improve heat-removal capability, but they also change the electrical composition and controllability of facility demand. Pumping power, coolant-temperature requirements, heat-rejection equipment, and thermal-storage opportunities become more important in the total energy-management problem. As a result, the flexibility of an AI data center cannot be assessed only from server power; the thermal response and cooling architecture must also be considered [18,19].

From the flexibility perspective, AI training and AI inference should be distinguished. Some training jobs can tolerate delayed start times, temporary interruption, checkpoint-based rescheduling, or migration among geographically distributed facilities. These characteristics make selected training workloads suitable for renewable-aware scheduling, carbon-aware operation, and demand-response programs. However, the available flexibility is reduced when training deadlines, expensive accelerator idle time, data locality, communication overhead, and synchronization among computing nodes are considered. Real-time inference generally provides less temporal flexibility because user-facing services are constrained by latency, availability, and response-quality requirements. Nevertheless, limited flexibility may still be obtained through server power capping, request routing, model placement, and coordination with UPS systems or dedicated storage.

Therefore, AI-driven data centers have a two-sided influence on renewable-rich power systems. Their large scale, rapid growth, high power density, and dynamic demand can intensify local capacity and balancing challenges. At the same time, selected training workloads, advanced cooling systems, and converter-interfaced storage resources may provide additional flexibility when they are coordinated carefully. The practical value of this flexibility depends on the proportion of training and inference workloads, accelerator utilization, service deadlines, cooling architecture, communication constraints, and the reserve margin required for reliable operation. AI data centers should consequently be modeled as heterogeneous electro-thermal-computing systems rather than as conventional constant-power loads.

2.3. Characteristics of High-Renewable-Penetration Power Systems

High-renewable-penetration power systems are usually marked by a large and still growing share of renewable generation, especially wind power and solar photovoltaic generation. The output of these sources strongly depends on weather conditions. Because of this feature, their power output is more changeable and less directly controllable than that of traditional synchronous generators [20,21]. In power-system research, these resources are often called variable renewable energy, or variable renewable energy (VRE). This name is used because their output depends on weather, sunlight conditions, and geographic distribution, rather than on planned fuel input [20,22]. As the share of VRE continues to rise, renewable generation is no longer just a small addition to the system. It is becoming a main factor that shapes system operation, reserve scheduling, and security assessment [20].

One key feature of high-renewable-penetration systems is generation variability, as shown in Figure 3. Wind power and solar power are very different from traditional thermal units. Their outputs change over many time scales, including seconds, minutes, hours, days, and even seasons [21,22]. This variability keeps changing the balance between generation and demand. It also changes the net demand that dispatchable resources need to supply.

Earlier studies have shown that renewable variability affects unit commitment and dispatch decisions. It also increases balancing needs and changes how operating reserves should be planned and used [21,23]. When renewable penetration becomes high, this variability can no longer be treated as a local problem. It becomes a system-wide operating feature that affects market scheduling, reserve allocation, and network use [20,22].

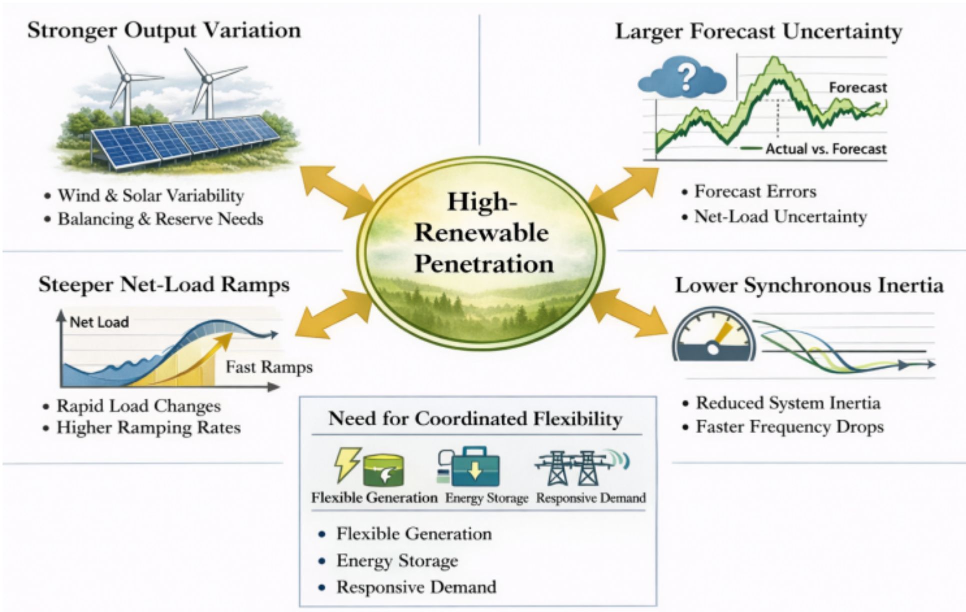

Figure 3. Key features of high-renewable-penetration systems.

Another important feature is forecast uncertainty. Forecast methods for wind and solar power have improved a lot in recent years. Even so, forecast errors still cannot be fully avoided, because renewable output is influenced by changing weather conditions and local climate behavior [23,24]. This problem is especially important because system operators do not only manage renewable output itself. They mainly manage net load, which is the difference between electricity demand and variable renewable generation. For this reason, errors in renewable generation forecasts directly become errors in net-load prediction. These errors then affect dispatch, reserve use, and real-time balancing [21,24]. Existing review papers have pointed out that better forecasting can lower operating costs and improve scheduling accuracy. At the same time, these studies also make it clear that uncertainty cannot be fully removed. It must be handled through system flexibility, reserve margins, and better coordination methods [23,24].

A third major feature is the stronger net-load ramping effect. In power systems with high shares of wind and solar generation, net load may change very quickly. This often happens when renewable output drops while demand is rising, or when renewable generation changes sharply in a short period [21,25]. This issue is especially clear in solar-rich systems. In such systems, net load may stay low at midday, but then rise steeply in the evening when solar output falls and demand remains high. Practical studies have shown that rising VRE penetration can greatly increase both the size and frequency of net-load ramping events. As a result, conventional generation, storage systems, interconnections, and flexible demand all need to respond faster and with larger capacity [25,26]. Therefore, the main operating challenge is not only to provide enough energy over the whole day. The system must also have enough ramping ability over short operating periods [21,26].

A fourth key feature is the decrease in system inertia. Traditional power systems mainly rely on synchronous generators. The rotating mass of these generators naturally provides inertia and helps slow down fast frequency changes after disturbances. However, as more of these generators are replaced by converter-based renewable resources, the total synchronous inertia of the system may drop a lot [27,28]. In low-inertia systems, frequency may change more quickly after sudden generation loss or load variation. This condition raises the risk of instability and makes fast control, fast reserve response, and advanced converter-based support more important [27,29]. Recent review studies have also pointed out that systems with high renewable penetration may face a larger rate of change in frequency, lower frequency nadirs, and stricter needs for primary frequency response and grid-forming control [28].

All of these features together lead to a much stronger need for flexibility. In this context, power-system flexibility means the ability of system resources to respond to expected or unexpected changes in supply and demand over different time scales, while still keeping operation secure and economical [30,31]. As renewable penetration increases, flexibility is needed not only for real-time balancing. It is also needed for unit commitment, reserve scheduling, congestion management, adequacy planning, and secure operation under system constraints. Review papers and international assessment reports have consistently shown that future flexibility cannot come from only one type of resource. It must come from a wider group of resources, including flexible generation, energy storage, transmission expansion, demand response, distributed energy resources, and digitally coordinated loads [29,32]. Therefore, compared with traditional grids dominated by conventional generation, high-renewable-penetration power systems are mainly marked by stronger output variation, larger forecast uncertainty, steeper net-load ramps, lower synchronous inertia, and a much greater need for coordinated flexibility in planning, scheduling, and real-time operation.

2.4. Coupling Challenges Between Data Centers and High-Renewable-Penetration Grids

The coupling of large-scale data centers with renewable-rich grids creates a new class of planning and operational tensions, as listed in Table 3. On one hand, data centers are increasingly viewed as promising flexible loads because of their large electricity demand, advanced control infrastructure, and growing use of on-site energy storage. On the other hand, their rapid load growth, high power density, and strict continuity requirements can intensify grid stress rather than alleviate it if the coupling is not properly coordinated. This duality becomes especially important in systems with high shares of wind and solar generation, where supply variability, limited local network capacity, and tighter reliability margins already make system operation more complex. Recent international assessments and reviews therefore indicate that the challenge is no longer simply how to power data centers with more renewable electricity, but how to coordinate data-center growth, grid expansion, and flexibility deployment in a way that preserves both grid reliability and computing-service continuity [1,8,10,17,33,34].

A first coupling challenge lies in the contradiction between high peak load concentration and local supply capability. Large-scale data centers are often connected as concentrated loads whose capacity can reach tens, hundreds, or even thousands of megawatts, which may exceed the pace at which substations, feeders, or regional transmission infrastructure can be reinforced [1,17,33,34]. NERC explicitly notes that emerging large loads, including data centers, can exhibit very rapid ramping characteristics and require carefully tailored interconnection and planning strategies; meanwhile, the latest reliability assessments also show that accelerated large-load additions are becoming a material driver of future adequacy and transmission concerns in several regions [17,33]. This issue becomes even more pronounced for AI inference-oriented and edge facilities, because they are often sited closer to end users for latency reasons and may therefore be connected to already constrained distribution infrastructure rather than to locations with abundant spare network capacity [34]. Consequently, even when sufficient bulk renewable energy exists at the regional level, local interconnection bottlenecks, feeder congestion, substation limitations, and upgrade lead times can delay or complicate data-center deployment and operation.

Table 3. Coupling challenges between data centers and high-renewable-penetration grid.

Category	Main Issue	Key Points	References
Overall coupling tension	Dual role of data centers	• Flexible load potential • Rapid load growth • High power density • Strict continuity needs • Coordination-dependent value	[1,8,10,17,33,34]
Planning challenge 1	Peak load vs. local supply	• Concentrated large loads • Slow grid reinforcement • Substation limits • Feeder congestion • Interconnection delays	[1,17,33,34]
Planning challenge 1	AI/edge siting pressure	• Latency-driven siting • Distribution-level constraints • Limited spare capacity • Local bottlenecks stronger	[17,33,34]
Operational challenge 2	Reliability vs. flexibility	• DR with constraints • SLA/QoS limits • Outage intolerance • Limited interruptibility	[8,10,35–37]
Operational challenge 2	Flexibility extraction limits	• Delay tolerance matters • Restart cost matters • Revenue sensitivity • Constrained optimization • Contract-based flexibility	[8,10,35–37]
Operational challenge 3	Renewable-workload mismatch	• Temporal mismatch • Spatial mismatch • Latency-sensitive tasks • Deadline constraints • User-demand coupling	[38–41]
Operational challenge 3	Workload shifting conditions	• Follow renewables • Delay flexible jobs • Inter-regional migration • Spare capacity needed • Accurate coordination	[39,40]
Operational challenge 3	Limits of shifting benefits	• Slow control response • Limited workload mobility • Weak spatiotemporal alignment • Network dependence	[40,41]
Cross-domain coupling	Multi-system coordination	• Power-computing coupling • Communication dependence • Scheduling interaction • Thermal constraints	[41]
Multi-timescale challenge	Different control horizons	• Long-term planning • Day-ahead scheduling • Real-time balancing • Workload dispatch; Cooling management	[10,40,41]

A second challenge is the conflict between supply reliability requirements and grid-oriented flexibility provision. In principle, data centers can support the grid through load shifting, backup-storage coordination, and participation in electricity-market services. However, their flexibility is constrained by strict quality of service (QoS) requirements, service level agreements (SLAs), power-quality limits, and outage-intolerance of critical digital services [8,10,35–37]. Recent reviews emphasize that although data centers have substantial theoretical flexibility potential, much of that flexibility cannot be treated as freely interruptible because computational services differ in delay tolerance, restart cost, and revenue sensitivity [8,10]. This is why many studies frame data-center participation in demand response (DR) as a constrained optimization problem rather than a simple curtailment problem: the facility may be able to move or defer some workloads, but only within operational windows that do not violate user commitments or degrade core services [35–37]. The proposed Green Service-Level Agreement and Green Service Delivery Agreement concepts in earlier work were introduced precisely to reconcile this tension, by explicitly embedding limited flexibility into service contracts rather than assuming that all data-center demand can be controlled like conventional flexible load [36]. Therefore, the key issue is not whether data centers can be flexible in principle, but how much flexibility can be extracted without undermining reliability, latency, throughput, and commercial service obligations.

A third challenge is the mismatch between renewable-generation variability and data-center workload rigidity. Renewable-rich grids are characterized by time-varying and location-dependent electricity availability, whereas a large share of data-center tasks remains latency-sensitive, deadline-constrained, or operationally coupled to user demand [38–41]. This mismatch appears in both temporal and spatial forms. Temporally, renewable output may be abundant at hours that do not coincide with computing demand or with the execution windows of critical tasks; spatially, renewable surpluses may occur in regions that are different from the locations where the most urgent computational workloads arise [38]. A large body of work on geographic and temporal workload shifting has therefore emerged to reduce this mismatch, showing that "follow-the-renewables" strategies, delayed execution of flexible jobs, and inter-regional workload migration can improve renewable utilization and reduce curtailment or emissions [39]. However, these studies also make clear that such strategies require sufficient spare computational capacity, acceptable delay tolerance, accurate coordination, and favorable network conditions. If control is too slow, if workload mobility is limited, or if local renewable conditions and computing demand are weakly aligned, the achievable benefit falls substantially [40]. In this sense, the renewable–data-center mismatch is not merely an energy-supply issue, but a coupled problem involving power systems, communication networks, computing schedules, and thermal-management constraints [41].

In addition to these three major contradictions, coupling challenges also arise from multi-timescale coordination. Power systems must coordinate long-term planning, day-ahead scheduling, and real-time balancing, whereas data-center operators must simultaneously manage infrastructure reliability, workload dispatch, cooling demand, and energy cost. Several recent studies suggest that local or short-horizon coordination alone is often insufficient, especially when data centers become large enough to influence regional power-system conditions or when renewable variability is strong [10,40]. As a result, future renewable-rich grids with large-scale data-center integration will require not only more flexible resources, but also more explicit coordination mechanisms between grid operators, market frameworks, and data-center energy-management systems [41]. Overall, the core coupling challenge is that data centers can potentially act as flexible grid-interactive assets, yet their physical scale, local network dependence, and stringent service constraints mean that this flexibility is conditional, limited, and highly dependent on where, when, and how the coupling is managed.

2.5. Conceptual Framework for Grid-Interactive Data Centers

As shown in Figure 4, a conceptual framework for grid-interactive data centers should clearly link the main internal subsystems of the facility with the external power system. In this way, the roles of flexibility, reliability, and operational coupling can be understood within one unified structure. In this framework, the data center should not be seen as a single passive load. Instead, it should be viewed as a layered cyber-physical energy system. This system includes an IT layer, a cooling and thermal management layer, a power continuity and storage layer, a backup power layer, and a supervisory energy management and control layer. All of these layers connect to the external grid through a point of interconnection and related power conditioning equipment [1,6,8,42]. This layered view is becoming more necessary. The reason is that large-scale data centers are growing quickly in electrical scale, facing stricter reliability requirements, and becoming more important in transmission and distribution planning because of their concentrated demand and fast-rising connection capacity [1,8,42]. In addition, the structure and performance requirements of data centers are still strongly influenced by infrastructure availability standards such as the Tier Classification System. This system links service continuity with redundancy, fault tolerance, and operational discipline [6].

The IT layer is placed at the center of this framework. This layer carries the main computing tasks of the facility and accounts for most of its energy consumption. It mainly includes servers, processors, accelerators, network devices, storage equipment, and different digital workloads. These workloads often have different requirements for latency, throughput, and interruption tolerance. In concept, this layer should be regarded as the main source of both electricity demand and operational rigidity, because different computing tasks do not have the same level of flexibility in time or location. The fast growth of artificial intelligence and high-performance computing workloads has further raised rack power density, electrical stress, and time-varying load behavior. As a result, the internal demand profile of data centers has become more dynamic than that of traditional cloud facilities [1,43,44]. At the same time, part of the IT workload can be scheduled, delayed, shifted across locations, or adjusted. Because of this feature, many recent studies treat the IT layer as a possible source of demand-side flexibility rather than only a fixed load [45–48]. Therefore, within this framework, the IT layer should be understood as both the main energy consumer and the first source of controllable flexibility. However, this flexibility is still limited by service and performance requirements [45–47].

The cooling and thermal management layer is closely connected to the IT layer. This layer includes computer room air conditioners, computer room air handlers, chillers, pumps, cooling towers, liquid cooling units, and related heat transfer infrastructure. This layer is essential because the heat generated by IT equipment must be removed continuously to maintain thermal safety, reliability, and performance. As computing density rises, the link between electricity demand and cooling demand becomes much stronger. Because of this trend, the total electrical behavior of the facility cannot be properly understood without considering the thermal subsystem [43,49]. In modern AI-oriented facilities, this issue is becoming even more important. High-density racks are pushing faster adoption of liquid cooling and other advanced thermal structures. These changes affect not only energy use, but also the dynamic response and controllability of the whole system [44,49]. From a conceptual view, the cooling layer acts as both a dependent load and a flexibility buffer. It follows changes in IT demand, but it can also provide some limited demand adjustment by using thermal inertia, chilled-water storage, or temperature tolerance margins [44,50].

Therefore, the framework should show a strong two-way link between the IT and cooling layers. Changes in computing intensity affect cooling demand, and cooling limits in turn shape the feasible operation of the IT layer.

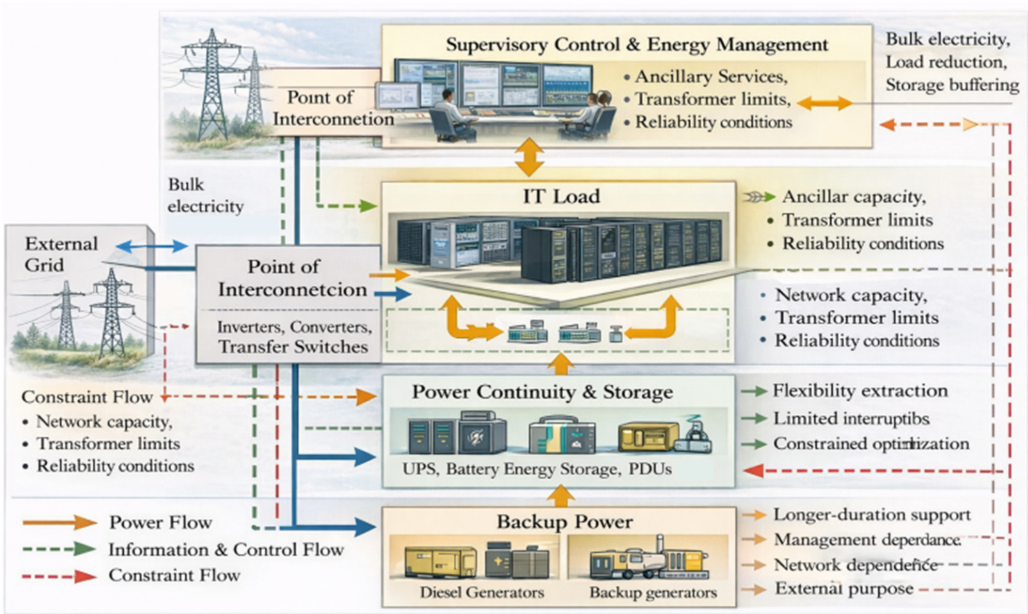

Figure 4. Conceptual framework for grid-interactive data centers [6,8,43,46,50].

The next important part of the framework is the power continuity and storage layer. This layer keeps power supply uninterrupted for critical loads and also supports fast electrical response and flexible operation. It usually includes UPS systems, battery energy storage systems, power distribution units, inverters, converters, and local bus structures that connect critical and noncritical loads [43,44]. In traditional data-center design, the main task of the UPS is to bridge short disturbances and maintain operation until backup generation starts. However, recent studies increasingly point out that UPS batteries and dedicated battery storage can do much more in grid-interactive operation. They can smooth power demand, support renewable integration, provide short-term reserve, and reduce stress at the grid connection point [44,51]. This change is especially important in AI data centers, where rapid load variation and high instantaneous demand may otherwise create serious power-quality and ramping problems [1,51]. For this reason, the proposed framework should treat the UPS and battery storage layer not only as a reliability asset, but also as a fast-response flexibility asset that connects internal demand dynamics with external grid conditions. In other words, this layer forms the main electrical buffer between mission-critical loads inside the data center and the changing conditions of the surrounding power system [44,51].

The backup power layer is used together with the storage layer. This layer usually includes diesel generators, gas generators, fuel cells, or hybrid standby resources. Its main traditional function is to keep the data center operating during long grid outages or serious power supply disturbances. However, in a grid-interactive framework, this layer also plays an important role in resilience, capacity support, and operational risk management [8,43,44]. In some new architectures, on-site generation is being evaluated together with storage and renewable resources as part of a coordinated energy portfolio, rather than only as an emergency subsystem [50,51]. This idea is especially important for mission-critical facilities that are connected to power grids with high renewable penetration but limited capacity. In these cases, backup resources are not only used to handle outages. They can also help reduce peak demand, support islanded operation, and provide more operating options when the grid is under stress [42,51]. Therefore, the backup power layer should be represented in the framework as a slower but longer-lasting support resource. It is different from the UPS and battery storage layer in both response speed and operational purpose.

To coordinate these internal layers, the framework also needs a supervisory energy management system, or a similar control architecture. This coordination layer exchanges information with IT scheduling systems, thermal controllers, storage controllers, backup power controllers, and external grid or market signals. Its task is to turn multiple goals into practical operating decisions across different time scales. These goals include reliability, cost, emissions, renewable use, and service quality [45–47,50]. In practice, this means the control layer must decide when computing workloads can be shifted, when cooling demand can be adjusted, when batteries should charge or discharge, when backup resources should stay idle or prepare for operation, and how the facility should respond to external price signals, congestion, or reliability requests [45,46,50]. Therefore, this layer is the main link between operational flexibility and operational constraints. Without this coordination layer, the data center is only a group of separate energy-using subsystems. With this layer, the facility can act as an integrated energy entity that interacts with the grid in an intelligent way.

The last part of the framework is the external grid interface. This interface connects the data center with the larger transmission or distribution system and determines how internal flexibility can affect external operation. Through this interface, the data center receives bulk electricity from the grid, may consume renewable-rich power, and in some cases may also exchange ancillary services or balancing support with the surrounding system [8,42,45,51]. From a conceptual point of view, three kinds of flows should be shown at this boundary. The first is power flow, which includes normal electricity import, peak reduction, and internal buffering through storage. The second is information and control flow, which includes market prices, demand response requests, reliability signals, and local operating data. The third is constraint flow. This means the grid places limits on the data center through feeder capacity, transformer ratings, interconnection rules, voltage quality, and reliability conditions. At the same time, the data center places counter-limits on the grid through its demand size, ramping behavior, and continuity requirements [8,42]. This two-way relationship is what gives real meaning to the term "grid-interactive". The data center does not simply take electricity from the grid. It develops and operates together with the grid.

Based on these parts, the conceptual framework can be summarized as a hierarchical and strongly coupled structure. The IT layer drives computing demand. The cooling layer turns IT heat into a linked electrical and thermal management problem. The UPS and battery storage layer provides fast continuity support and short-term flexibility. The backup power layer provides longer-duration resilience. The energy management system coordinates all of these internal resources. The external grid interface then connects internal operation with system-level energy, market, and reliability conditions [6,43–49,51]. Within this structure, several interactions are especially important. First, there is the IT-cooling coupling, because computing intensity determines thermal demand. Second, there is the coupling between IT and cooling on one side and UPS and battery storage on the other side, because storage helps buffer fast internal fluctuations. Third, there is the coupling between UPS and battery storage and backup power, because short-term and long-term continuity resources must be arranged in the right order. Fourth, there is the coupling between the internal system and the grid, because grid conditions affect operating decisions, while internal resource actions can also affect external ramps, peaks, and reserve needs [50]. This framework provides the basis for the following sections of this review. It allows the available flexibility resources, grid interaction mechanisms, and operational impacts to be discussed in a clearer and more comparable way.

3. Flexible Resources for Large-Scale Data Centers

Large-scale data centers contain multiple forms of operational flexibility that can potentially support power systems with high renewable penetration. Unlike traditional electricity users, data centers are not characterized by a single fixed demand profile. Their flexibility can arise from controllable computing workloads, thermally buffered cooling systems, on-site UPS and battery storage, and backup generation resources. These resources differ significantly in response speed, controllability, duration, operational purpose, and service constraints. Some of them can provide very fast electrical support, while others are more suitable for slower load shaping or longer-duration balancing. Therefore, a systematic review of data-center flexibility should not treat these resources as a uniform category. Instead, they should be understood as a multi-layer flexibility portfolio with complementary characteristics across different time scales. This section reviews the major flexible resources available in large-scale data centers and discusses their operating principles, flexibility mechanisms, and coordination potential under renewable-rich grid conditions.

3.1. IT and Workload Flexibility

IT and workload flexibility is one of the most important flexible resources in large-scale data centers. Traditional industrial loads usually follow fixed operating patterns. In contrast, part of the computing demand in data centers depends on how digital tasks are arranged, delayed, assigned, or moved across servers and different sites. This feature shows that the electricity demand of a data center is not always completely fixed. Instead, operators can adjust some workloads in time or place so that demand can better fit the condition of the power system [45,46]. Researchers have discussed this idea for more than ten years. In earlier studies on cloud-grid coordination and geographically distributed data-center operation, researchers already treated computing demand as a schedulable load rather than a completely passive load, as shown in Figure 5.

The main reason why workload flexibility exists is that workloads are not all the same. In real operation, different computing tasks have different needs in delay tolerance, interruption tolerance, and location dependence. Interactive services and delay-sensitive applications usually leave very little room for adjustment. By contrast, batch computing tasks, background data analysis, and some artificial intelligence (AI) training jobs can often be shifted within an allowed time range [47,52]. This difference makes it possible for workload management to change the shape of power demand without directly affecting critical services. In general, workload flexibility mainly appears in two forms. The first form is temporal shifting. This means that operators delay or advance selected tasks within an acceptable execution period. The second form is spatial shifting. This is also called geographical load balancing. In this case, operators move workloads among different data-center sites according to electricity prices, renewable energy availability, carbon intensity, or local grid conditions [53].

From the viewpoint of the power system, workload flexibility is valuable because it can change when and where electricity is used. This resource is different from a battery or a generator. It does not send power back to the grid directly. Instead, it provides support by changing the demand profile over timescales from minutes to hours and from hours to days [54]. Earlier studies showed that workload shifting can reduce simultaneous peak demand and lower operating costs, especially when it is used together with local generation or other on-site resources. More recent studies have further connected workload scheduling with carbon-aware operation. These studies showed that operators can move flexible computing tasks to periods with lower carbon intensity or stronger renewable generation [55]. For this reason, workload flexibility is especially important in power systems with high renewable penetration. In such systems, net-load fluctuations and regional differences in renewable output create a strong need for demand that can be adjusted in a controlled way.

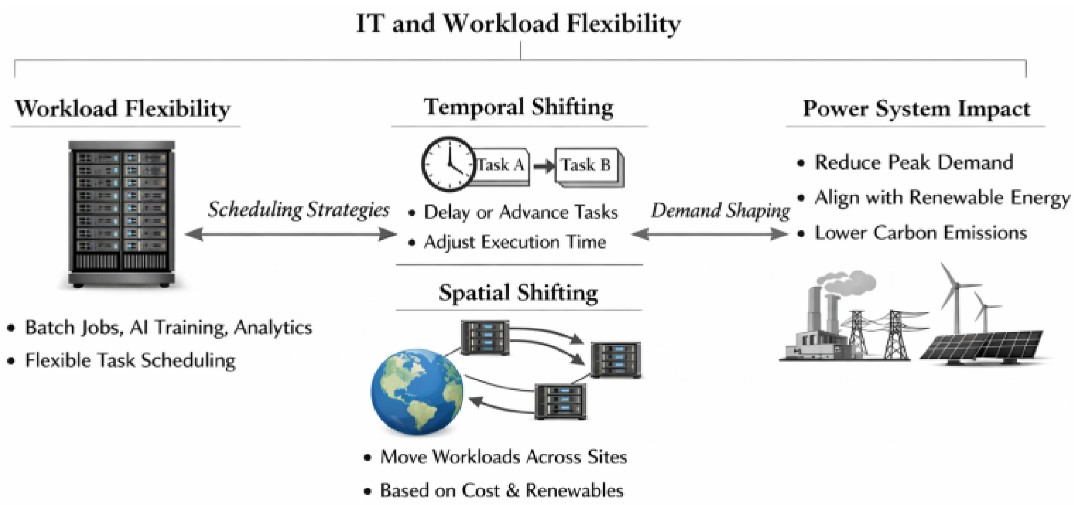

Figure 5. IT and workload flexibility pathways for grid-interactive data centers.

However, the amount of flexibility that operators can obtain from IT workloads is still limited by many practical constraints [48]. This point is very important in the existing literature. Data-center operators must meet SLAs, quality of service (QoS) requirements, delay limits, throughput goals, and reliability demands. At the same time, spatial migration also needs spare computing capacity, acceptable communication delay, and enough data-transfer ability between sites. Therefore, workload flexibility should not be seen as a demand resource that can be dispatched freely at any time. Its actual effect depends on workload type, scheduling freedom, infrastructure backup, and system coordination ability. In general, IT and workload flexibility should be viewed as a conditional resource that is restricted by computing tasks and operating requirements. When operators respect these constraints, this resource can still provide meaningful support for renewable-rich power systems.

Although IT and workload flexibility have been widely identified as a promising demand-side resource, the reported benefits should be interpreted with caution. First, many studies assume that a sufficiently large proportion of workloads is delay-tolerant or geographically movable, whereas real data-center portfolios often contain a substantial share of latency-sensitive and location-dependent services. Second, simulation-based studies commonly assume accurate workload forecasts, sufficient spare computing capacity, and negligible migration overhead. These assumptions may overestimate the flexibility available in real operation, particularly when data transfer, cybersecurity, and network congestion are considered. Third, the benefits of temporal and spatial workload shifting are not always consistent across studies because they depend strongly on workload composition, regional electricity prices, carbon intensity, communication conditions, and the degree of renewable-energy correlation among different sites. Therefore, workload flexibility is most applicable to batch processing, background analytics, and selected AI-training tasks, but is less suitable for real-time inference, financial transactions, and other latency-critical services. Its practical value should consequently be assessed as conditional and service-specific rather than as a uniformly dispatchable resource.

AI workloads require a more detailed flexibility classification than conventional cloud workloads. Large-scale training tasks may offer temporal flexibility through delayed execution, checkpointing, interruption, and restart, and may also provide spatial flexibility when datasets, accelerators, and communication capacity are available at multiple sites. However, the energy cost of restarting, accelerator underutilization, synchronization overhead, and training deadlines can substantially reduce this theoretical potential. In contrast, real-time inference is generally latency-sensitive and closely coupled to user demand, so direct workload shifting is more limited. For inference-oriented facilities, short-term flexibility is more likely to come from request routing, server power capping, model placement, and coordination with UPS or BESS resources than from long-duration task postponement. Therefore, the flexibility of an AI-driven data center should be evaluated according to the specific proportions and operating characteristics of training and inference workloads.

3.2. Cooling and Thermal Flexibility

Cooling and thermal flexibility is another important flexible resource in large-scale data centers. This resource is different from IT loads. IT loads are mainly limited by computing-service needs, while cooling systems can often offer short-term flexibility by using thermal inertia and allowable temperature margins. In modern data centers, the cooling system usually includes computer room air conditioners (CRACs), computer room air handlers (CRAHs), chillers, pumps, cooling towers, liquid-cooling units, and chilled-water loops. Server power is finally converted into heat, so cooling demand is closely linked to IT demand. As rack power density keeps rising, especially in artificial intelligence (AI) and HPC facilities, this link between electricity use and heat becomes even stronger. As a result, cooling systems now account for a large part of total electricity use in many facilities [7,43,44,49].

The main advantage of cooling flexibility is that thermal states do not change immediately. This feature means that operators can adjust cooling demand for a short period without directly crossing equipment safety limits. In practice, operators can create this flexibility in several ways. They can adjust supply-air temperature, reset chilled-water temperature, control variable-speed fans and pumps, improve air-flow organization, use free cooling, apply water-side economizers, and coordinate different cooling devices together [56–58]. In addition, studies on thermal-aware scheduling show that workload placement and cooling control should not be treated as two separate problems. Server utilization, air distribution, and room thermal behavior affect each other directly [47,59,60]. Therefore, cooling flexibility is not only a matter of mechanical equipment. It is a coupled thermal-management resource, and its performance depends on both system design and IT operating conditions.

From the time-scale perspective, cooling and thermal flexibility is mainly useful from minutes to hours. This type of flexibility is suitable for peak shaving, short-term load smoothing, and partial matching with variable renewable energy output. Compared with direct workload reduction, cooling-side adjustment can provide flexibility without immediately lowering computing throughput. This feature makes it attractive for demand response and renewable-following operation. However, its usable range is still limited. Data centers must stay within the temperature and humidity ranges required for reliable IT operation. If operators reduce cooling too aggressively, the system may face hot spots, local air recirculation, or delayed thermal rebound [47,49]. In addition, air-cooled systems and liquid-cooled systems do not provide the same type of flexibility. Recent review studies show that high-density AI facilities are moving more and more toward liquid cooling. This trend improves heat removal capability, but it also changes the control features and design limits of the thermal system [61,62].

Another important approach is the use of thermal energy storage (TES). This method allows operators to store cooling capacity in advance and use it later when needed. In this way, part of the cooling load can be separated from real-time electricity demand. This feature is useful when operators want to move cooling consumption away from high-price periods or shift it toward hours with more renewable energy supply. Existing studies show that chilled-water storage, ice storage, and phase-change-material-based storage can enlarge the flexibility range of cooling systems and improve both energy efficiency and system resilience [50]. Overall, as shown in Table 4, cooling and thermal flexibility should be understood as a resource that is limited by thermal conditions but is still highly practical. Its main strength is that it can reshape facility demand without directly interrupting core digital services. Its main weakness is that the adjustment range is limited, and its actual value depends strongly on thermal safety limits, system structure, and coordinated control ability.

Table 4. Summary of cooling and thermal flexibility in large-scale data centers.

Aspects	Key Points	References
Physical basis	• Cooling-coupled load • Thermal inertia use • Density-driven growth	[7,43,44,49]
Main flexibility forms	• Temperature reset • Flow-rate control • Airflow management • Free cooling use	[56–58]
Coordination mechanisms	• IT-cooling coupling • Joint scheduling • Thermal-aware control	[47,59,60]
Grid-side value	• Peak shaving • Load smoothing • Renewable alignment • Demand response support	[47,49]
Constraints and extensions	• Thermal safety limits • Hotspot risk • Rebound effect • Thermal storage support	[50,61,62]

The practical value of cooling flexibility varies considerably among data-center configurations. Air-cooled facilities usually provide relatively simple temperature-reset and airflow-control opportunities, but their flexibility is limited by hotspot formation, air recirculation, and delayed thermal rebound. Liquid-cooled facilities can support higher rack densities and more effective heat removal, but their control dynamics, pumping requirements, and allowable operating margins differ from those of conventional air-cooled systems. In addition, many optimization studies represent the thermal subsystem using simplified lumped models, which may not capture local temperature gradients and equipment-level thermal risks. Consequently, reported peak-shaving benefits obtained from simplified simulations may be difficult to reproduce in heterogeneous facilities. Cooling flexibility is therefore more suitable for short-duration load shaping than for sustained power reduction, unless dedicated thermal energy storage is available. Its implementation also requires facility-specific thermal models and conservative safety margins.

3.3. UPS, BESS, and Backup Power Resources

Uninterruptible power supply (UPS) systems, battery energy storage systems (BESSs), and backup power units are the main electrically controllable flexible resources in large-scale data centers, as shown in Figure 6. In traditional data-center structures, operators install these assets mainly to maintain continuous service during grid disturbances, voltage drops, and power outages. However, under grid conditions with high renewable penetration, these assets are no longer seen only as reliability support equipment. Researchers and operators now also view them as flexible electrical resources that can help with peak shaving, short-term balancing, renewable-energy integration, and resilience improvement [8,42]. Compared with workload flexibility and cooling flexibility, these resources are connected more directly to the grid interface. They also usually offer faster and more accurate control performance [8,63].

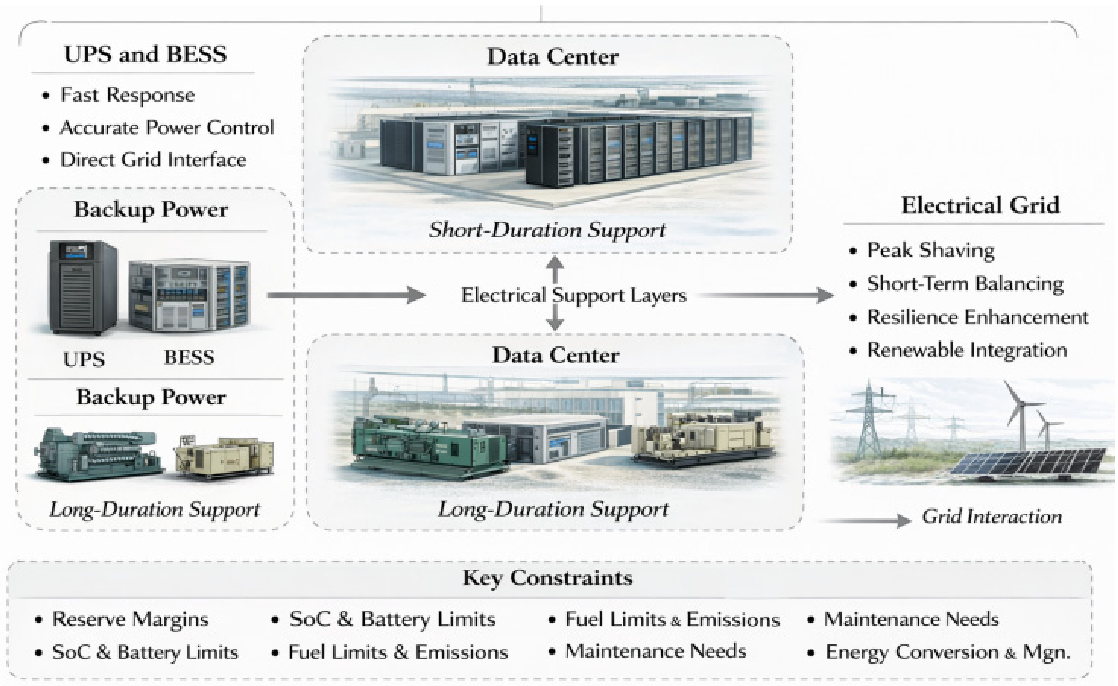

Figure 6. UPS, BESS, and backup power resources in large-scale data centers.

Among these resources, UPS systems and battery-based storage are especially important for fast-response operation. Existing studies show that batteries already deployed for backup purposes in data centers can also support demand response, peak reduction, and short-duration ancillary services under proper control strategies. This can be done without changing the fundamental role of the data center as a mission-critical facility [54,63,64]. More recent studies further show that coordinated control of multiple UPS batteries, as well as the use of storage-enhanced UPS structures, can expand the available flexibility of data centers. At the same time, these methods can improve renewable-energy use and reduce operating costs [51,65]. Therefore, UPS systems and BESSs should be understood as the short-duration and high-speed flexibility layer inside the data center. Their main advantages are fast response, accurate power regulation, and direct compatibility with converter-based power systems [42,65,66].

Backup power resources play another role, but this role is also very important. These resources usually include diesel generators, natural-gas generators, fuel cells, and hybrid standby systems. In the past, operators mainly reserved these units for rare but serious outage events. However, recent studies show that backup resources can also support microgrid operation, local adequacy, peak reduction, and longer-duration resilience when they are coordinated properly with storage systems and supervisory control [67–69]. This role is becoming more important because many data centers now face grid-connection delays, local network limits, and growing pressure to keep services running under stressed grid conditions. In this setting, backup power should be viewed as a slower but longer-lasting support layer that works together with UPS systems and battery storage [42,67,68].

Even though these resources have clear flexibility potential, their real use is still limited by many practical factors. For battery-based resources, the available flexibility depends on the state of charge (SoC), the reserve margin kept for emergency backup, battery aging caused by charge and discharge cycles, and the design principles of the data-center tier and internal power structure [54,64,70]. Recent reliability studies show that using backup batteries for flexible operation can be economically attractive. However, the feasible level depends strongly on reliability goals, local grid conditions, and the need to keep enough backup energy available [64,70,71]. Backup generators face other limits as well. These limits include fuel supply, startup delay, maintenance needs, emissions, and regulatory requirements. As a result, operators cannot treat backup generators as completely free and fully available flexibility resources [68,69].

Another important trend is that electrical support resources are becoming part of more coordinated data-center energy systems. Recent reviews point out that future facilities, especially AI-oriented data centers, will likely rely more on hybrid storage, advanced battery management systems (BMSs), and coordinated operation of UPS systems, storage units, and distributed on-site generation [13,72,73]. At the same time, optimal dispatch studies show that these resources can create significant economic value when operators use them for smart-grid services. However, this value is highly affected by market conditions, battery cost, lifecycle aging, and continued load growth [51,73,74]. For this reason, the practical value of these resources does not come only from their individual performance. Their value also depends on how well operators coordinate them across different time scales and operating goals [8,13,74].

Overall, UPS systems, BESSs, and backup power resources form the core electrical flexibility portfolio of large-scale data centers. UPS systems and batteries mainly provide very fast and accurate short-term support. Backup generation mainly provides slower but longer-duration support for resilience. Their combined role becomes especially important in renewable-rich power systems, where short-term fluctuations, low inertia, and supply uncertainty all increase the need for flexible and reliable electrical support [8,42,66]. However, these resources should not be understood as grid assets without limits. Their real flexibility depends on emergency backup requirements, degradation limits, operating cost, environmental constraints, and the quality of coordination within the wider data-center energy management framework [64,72,74].

UPS systems, dedicated BESSs, and backup generators provide more direct electrical controllability than workload and cooling resources, but their grid-support value should not be evaluated only according to response speed and power capacity. Dual use of UPS batteries for reliability backup and grid services creates an inherent conflict between economic utilization and emergency readiness. Frequent cycling may accelerate battery degradation, while maintaining a high reserve margin reduces the energy available for market participation. The conclusions reported in techno-economic studies are therefore sensitive to battery replacement cost, degradation models, electricity-market prices, outage probabilities, and required backup duration. Backup generators provide longer-duration support, but diesel-based operation may increase local emissions and may be restricted by environmental regulations, fuel availability, maintenance requirements, and startup delay. For these reasons, BESS-based flexibility is more suitable for fast and short-duration services, whereas backup generation is primarily appropriate for resilience and emergency support rather than frequent market dispatch. Studies that ignore degradation, emissions, and reserve-preservation constraints may overestimate the long-term flexibility and profitability of these resources.

3.4. Integrated Flexibility Potential and Resource Coordination

The flexibility of large-scale data centers should be understood as an integrated resource portfolio rather than a group of separate resources. IT workload flexibility, cooling and thermal flexibility, UPS systems, BESSs, and backup power units are not the same in response speed, support duration, controllability, or operating limits. Recent review studies show that the overall value of data centers becomes much higher when operators coordinate these different resources across multiple operating time scales instead of using them one by one in an isolated way [8].

To improve the comparability of the reviewed flexibility resources, Table 5 summarizes their typical response times, support durations, adjustable quantities, grid-service applications, principal quantitative constraints, and evidence maturity. The values are reported as representative ranges or orders of magnitude because the available flexibility varies considerably with facility architecture, workload composition, thermal design, storage sizing, and backup requirements. This comparison shows that electrical storage resources generally provide the fastest response, while workload and thermal resources offer longer but more condition-dependent flexibility.

The main reason is that these resources can complement each other. Workload-side flexibility mainly supports adjustment from minutes to hours and from hours to days through temporal scheduling and geographical load balancing [46,54]. Cooling-side flexibility mainly supports short-term demand shaping through thermal inertia and coordinated cooling–workload management [47,60]. In contrast, UPS systems and BESSs can provide very fast electrical response, while backup generation and microgrid-oriented structures can provide slower but longer-lasting support [65,67]. Therefore, no single resource can meet all grid-support needs on its own. In practice, useful flexibility depends on the coordinated use of fast electrical buffering, thermal adjustment, and slower workload rescheduling [8,46,47,54,60,65,67].

This feature also means that coordination should be carried out across multiple time scales. Batteries and UPS systems are more suitable for fast balancing and ramp smoothing. Cooling systems are more suitable for intermediate peak shaving. Workload scheduling is more suitable for longer-term adjustment related to cost, carbon, and renewable-energy conditions [46,47,54,60,65]. Studies on carbon-aware computing further show that effective coordination should not look only at electricity demand. Operators should also consider carbon intensity, workload deadlines, storage condition, and reliability limits together [48].

A comparison of the reviewed resource categories reveals that no flexibility option is dominant across all performance dimensions. UPS systems and BESSs offer the fastest and most accurate electrical response, but their duration is limited and their repeated use is constrained by degradation and backup-reserve requirements. Cooling systems provide relatively low-disruption flexibility because they do not directly curtail computing services, but their usable range is facility-specific and limited by thermal safety and rebound effects. Workload shifting can provide longer-duration and geographically distributed flexibility without battery degradation, but it is highly dependent on workload characteristics, communication infrastructure, and SLA constraints. Backup generation can sustain operation for a longer period, but it has slower dynamics and may introduce fuel, emission, and regulatory concerns. Therefore, the suitability of each resource depends on the required response time, service duration, reliability margin, and operational objective.

https://doi.org/10.3390/en19153555

Table 5. Quantitative and operational comparison of major flexibility resources in large-scale data centers.

Flexibility Resource	Typical Response Time	Typical Support Duration	Main Adjustable Quantity	Typical Grid-Support Function	Main Quantitative Constraint	Evidence Level
Temporal workload shifting	Minutes to hours	Hours to days	Delay-tolerant IT load	Peak shaving, renewable following, energy arbitrage	Task deadlines, allowable delay, flexible workload share	Mainly simulation and optimization; limited operational evidence
Spatial workload migration	Minutes to hours	Hours to days	Computing load transferred among sites	Regional congestion relief, carbon-aware operation	Network latency, migration time, bandwidth, spare computing capacity	Mainly simulation and case studies
Cooling-system control	Seconds to minutes	Several minutes to several hours	Chiller, fan, pump, and cooling power	Load smoothing, peak shaving, short-term demand response	Temperature and humidity limits, hotspot risk, rebound effect	Simulation, laboratory studies, and limited pilots
Thermal energy storage	Minutes	Several hours	Stored chilled water, ice, or phase-change cooling capacity	Peak shifting and renewable-energy alignment	Storage capacity, charging/discharging efficiency, thermal losses	Simulation and engineering applications
UPS batteries	Milliseconds to seconds	Seconds to tens of minutes	Active power at the grid interface	Fast frequency response, regulation, ramp smoothing	Minimum state of charge, emergency reserve, cycle degradation	Laboratory, pilot, and limited operational evidence
Dedicated BESS	Milliseconds to seconds	Minutes to several hours	Charge/discharge power	Frequency regulation, reserve, peak shaving, energy shifting	Energy capacity, SoC, degradation, converter rating	Commercially mature, but data-center-specific use remains limited
Backup generators	Seconds to minutes	Hours to days, subject to fuel supply	On-site generation output	Emergency supply, resilience, capacity support	Startup delay, fuel availability, emissions, maintenance	Extensive emergency operation; limited routine grid-service use

Note: The reported response-time and duration ranges represent typical orders of magnitude identified across the reviewed literature rather than universal technical limits. The actual values depend on facility architecture, workload composition, thermal design, storage sizing, control strategy, and reliability requirements.

The reviewed literature also contains differences in the estimated flexibility potential of data centers. Some studies report substantial cost, emission, and peak-demand reductions, whereas others identify only limited flexibility after reliability and service constraints are included. These differences do not necessarily represent direct contradictions. In many cases, they arise from different assumptions regarding flexible workload share, battery reserve margin, migration cost, thermal limits, renewable availability, market prices, and system boundaries. Studies based on idealized simulations generally report larger benefits than studies that incorporate detailed network, thermal, degradation, or SLA constraints. Accordingly, future comparisons should use transparent assumptions and standardized performance indicators, and should distinguish theoretical potential from technically feasible and operationally verified flexibility.

The coordinated use of workload flexibility, cooling systems, UPS batteries, BESSs, and backup power requires an appropriate energy management framework. In general, the main decision variables include workload execution time and location, cooling set points, battery charging and discharging power, UPS reserve margins, and backup-resource scheduling. The optimization objectives usually include electricity-cost reduction, peak shaving, renewable-energy utilization, carbon-emission reduction, and grid-service provision. At the same time, the decisions must satisfy SLA and QoS requirements, thermal safety limits, battery state-of-charge constraints, degradation limits, backup-duration requirements, and grid-connection limits.

Existing studies mainly use deterministic optimization, stochastic or robust optimization, model predictive control, and multi-objective or hierarchical energy management. Deterministic programming is transparent and suitable for day-ahead scheduling, but its performance may deteriorate when workload, renewable generation, or electricity prices deviate from forecasts. Stochastic and robust methods can address uncertainty, although stochastic methods require reliable probability information and robust methods may produce conservative schedules. Model predictive control is more suitable for real-time coordination because it can update decisions using new measurements and forecasts, but its performance depends on model accuracy and computational speed. Multi-objective and hierarchical methods are useful for balancing cost, carbon emissions, reliability, service quality, and battery degradation across different time scales. Therefore, optimization-method selection should depend on the operating horizon, uncertainty level, reliability requirement, and available computational capability.

Table 6 shows that no single optimization method is suitable for all operating conditions. Day-ahead scheduling, real-time control, uncertainty management, and reliability protection usually require a combination of different methods.

For AI-driven data centers, multi-timescale coordination is especially important because training workloads, inference services, high-density cooling systems, and fast electrical storage exhibit substantially different response characteristics. Training rescheduling is more suitable for hour-level energy and carbon optimization, whereas UPS systems, BESSs, server power capping, and cooling control are more suitable for second-to-minute-level grid support.

The maturity of the available evidence also differs among flexibility resources. Studies on workload shifting, workload migration, and cooling flexibility are still dominated by theoretical analysis and simulation, although selected industrial applications have demonstrated the feasibility of carbon-aware scheduling and cooling management. UPS-based flexibility has progressed from simulation and hardware validation to pilot projects and limited operational grid-service deployment. Dedicated BESSs have extensive operational experience in the broader power sector, but publicly available evidence specifically related to data-center participation remains limited. Backup generators are mature for emergency supply, whereas their routine application for grid services is constrained by fuel consumption, emissions, maintenance requirements, and regulatory restrictions. Therefore, theoretical or simulation-based flexibility potential should not be regarded as equivalent to sustained operational capability.

Table 6. Comparison of optimization and energy management methods for grid-interactive data centers.

Method	Main Application	Main Advantage	Main Limitation
Deterministic optimization	Day-ahead workload, cooling, and storage scheduling	Clear formulation and interpretable results	Sensitive to forecast errors
Stochastic optimization	Scheduling under renewable, workload, and price uncertainty	Represents probabilistic uncertainty	Requires probability distributions and many scenarios
Robust optimization	Reliability-oriented scheduling under uncertainty	Maintains feasibility under adverse conditions	May be conservative
Model predictive control	Real-time and rolling-horizon coordination	Corrects forecasts and operating deviations	Depends on model accuracy and computation time
Multi-objective/ hierarchical optimization	Coordination of cost, carbon, reliability, and QoS across time scales	Handles conflicting objectives and different resource dynamics	Requires careful weighting and inter-layer coordination

4. Grid Interactions of Large-Scale Data Centers

Large-scale data centers interact with power systems not only as high-density electricity users, but also as emerging flexible resources with increasing internal controllability. In power systems with high renewable penetration, this interaction is becoming more important. The reason is that data centers can affect system operation in several ways. They can reshape electricity demand, carry out renewable-aware scheduling, use on-site storage and UPS systems for support, and in some cases participate in grid services. However, the actual value of this interaction depends on several conditions. Operators need to coordinate different internal resources effectively. They also need to ensure that grid-support actions do not violate reliability requirements, SLAs, or computing-performance limits. Therefore, this section reviews the main pathways through which large-scale data centers interact with power systems. The discussion focuses on three aspects: flexible load behavior, participation in electricity markets and grid services, and the emerging grid-support role of converter-based resources.

4.1. Data Centers as Large Flexible Loads

Large-scale data centers can no longer be seen only as passive electricity users. From the viewpoint of power-system operation, researchers increasingly regard them as large flexible loads. The reason is that part of their electricity demand can be adjusted through workload scheduling, spatial load redistribution, and coordinated control of internal resources. This feature is different from that of many traditional industrial loads. In many industrial systems, power consumption is closely linked to physical production processes, so the room for adjustment is often limited. Early studies on cloud–grid coordination had already pointed out this difference, as shown in Table 7. These studies showed that data centers can respond to grid conditions by adjusting computing demand, rather than only taking electricity as an outside input that cannot be changed. This finding established the basic view that data centers can become controllable demand-side resources in future smart grids [45–47,75].

Table 7. Summary of data centers as large flexible loads.

Aspects	Key Points	References
Fundamental role	• Large flexible load • Controllable demand-side resource • Different from industry loads	[45–47,75]
Main flexibility forms	• Temporal shifting • Spatial shifting • Workload coordination • On-site energy coupling	[48,53–55]
Grid-side value	• Peak reduction • Renewable matching • Carbon-aware operation • Location-based adjustment	[48,53–55]
Main constraints	• SLA constraints • QoS limits • Thermal safety	[42,76–78]

As large flexible loads, data centers mainly interact with the grid by changing the timing and location of electricity demand. Temporal flexibility comes from the fact that some computing tasks can tolerate delay and can be moved within an acceptable execution period. Spatial flexibility comes from geographical load balancing across different sites with different electricity prices, renewable-energy availability, or carbon intensity. In addition, some facilities can combine workload control with on-site energy management, and this further improves their ability to reshape demand. These mechanisms are valuable in renewable-rich power systems because they can reduce coincident peaks, improve the match between demand and renewable supply, and move electricity use toward periods or locations with better system conditions. More recent studies on carbon-aware scheduling further show that operators can align temporally flexible computing tasks with periods of lower-carbon electricity. This trend further strengthens the role of data centers as responsive flexible loads with growing policy relevance [48,53–55].

However, treating data centers as flexible loads does not mean that operators can dispatch their demand freely. In practice, the available flexibility is strongly limited by internal service duties and uncertain operating conditions. Data-center operators must meet SLAs, QoS requirements, thermal safety limits, and reliability targets. For this reason, some studies describe data-center participation in DR programs as a constrained optimization problem rather than a simple load curtailment problem. Existing studies have examined smart DR participation, reliability-aware aggregation, and QoS-guaranteed reserve provision. These studies show that operators must balance flexibility use against penalty costs, uncertain workload arrivals, and user-performance requirements. Therefore, the actual flexibility of data centers depends not only on their electrical scale, but also on workload type, control structure, and contractual risk [42,76–78].

Overall, the importance of data centers as large flexible loads comes from the combination of large scale, controllability, and growing internal energy resources. Unlike small responsive loads, large-scale data centers can influence local peak demand, feeder loading, and system balancing in a visible way. Unlike purely interruptible loads, they offer several ways to reshape demand, including computation-aware scheduling, geographically distributed operation, and coordination with on-site storage or UPS infrastructure. Recent review studies therefore argue that the power-system value of data centers becomes highest when they are understood not only as a source of load growth, but also as emerging flexibility assets. Under proper coordination, their participation can support reliability, renewable-energy integration, and more adaptive grid operation [8].

4.2. Participation in Grid Services and Electricity Markets

Large-scale data centers can participate in grid services and electricity markets through controllable demand, on-site storage, UPS systems, and backup power resources. From the viewpoint of power-system operation, this participation is attractive because data centers combine large electrical capacity with relatively advanced monitoring and control systems. As shown in Table 8, recent review and survey studies show that data centers are no longer considered only for DR. Researchers now also discuss their broader market roles, including energy-market optimization, reserve provision, and ancillary-service support. This change is especially important in power systems with high renewable penetration. In these systems, market value is becoming more closely related to flexibility, ramping capability, and fast response, rather than to electricity consumption alone [8,35,79,80].

Table 8. Summary of data center participation in grid services and electricity markets.

Aspects	Key Points	References
Participation foundation	• Large controllable load • Internal flexibility resources • Advanced control infrastructure	[45–47,75]
Grid service forms	• Demand response • Energy market participation • Reserve provision • Ancillary services	[46–48,53–55,75]
Technical realization	• Workload reshaping • Server power capping • Aggregated flexibility • QoS-aware control	[48,53–55]
Storage-based support	• UPS-assisted response • BESS-based support • Fast power control • Grid-interactive design	[76]
Constraints and market dependence	• Reliability constraints • Battery degradation • Qualification rules • Revenue uncertainty	[8,42,77,78]

In practice, one important participation pathway is the provision of balancing and ancillary services. Earlier studies showed that data centers can provide short-term load adjustment and reserve-like response by reshaping computing demand, limiting server power, or coordinating on-site energy resources. Simulation and optimization results indicate that this type of participation can reduce electricity cost or create extra revenue, while causing only limited influence on QoS when the control strategy is designed well. More recent studies further suggest that aggregation, reliability-aware scheduling, and regulation service reserve (RSR) control can improve the feasibility of market participation. These methods can reduce uncertainty and help operators maintain service commitments. Therefore, from a technical point of view, data centers do not only respond passively to electricity prices. Under suitable constraints, they can also act as dispatchable flexible loads in ancillary-service settings [35,79–81].

Another important pathway is participation through storage-based and grid-interactive structures. In this case, the data center does not rely only on workload flexibility. It also uses battery-backed UPS systems or dedicated BESSs to respond to market signals and system needs. This approach expands the possible service range. It allows the data center to move from slower demand reshaping to faster active-power response, including short-duration reserve, frequency-related support, and dynamic grid balancing. Industry reports and review studies indicate that these structures are becoming more important as converter-based power systems place greater value on speed and controllability. At the same time, the practical value of these services depends strongly on how operators manage internal reserve margins, battery aging, and mission-critical reliability requirements [82].

However, the actual market participation of large-scale data centers is not decided by technical ability alone. Market rules and regulatory frameworks also play an important role. In the United States, Federal Energy Regulatory Commission (FERC) Order No. 841 required organized wholesale markets to remove barriers to the participation of electric storage in energy, capacity, and ancillary-service markets. This change is directly relevant to storage-enabled data-center participation. At the regional level, market rules documented by PJM Interconnection (PJM) for energy and ancillary services, and by the Electric Reliability Council of Texas (ERCOT) for load-resource participation, show that flexible demand and storage can enter market-based service arrangements when they meet defined qualification and performance requirements. Overall, large-scale data centers clearly have the potential to participate in grid services and electricity markets. However, their actual role depends on the match between internal flexibility, reliability constraints, and external market design [42,77,83–87].

Beyond conventional energy and demand-response programs, data centers may also participate in emerging flexibility and capacity markets. In flexibility markets, system operators or distribution utilities procure adjustable demand to manage short-term balancing needs, renewable-energy variability, congestion, and local network constraints. Data centers can provide such flexibility through workload rescheduling, cooling adjustment, UPS batteries, and dedicated BESSs. In capacity markets, these resources may contribute by reducing demand or supplying stored energy during system-stress periods. However, participation generally requires verified availability, minimum response duration, telemetry, baseline calculation, and compliance with performance obligations. Failure to deliver the committed capacity may also expose operators to financial penalties [42,77,83–86].

Local energy markets and transactive energy systems provide another possible participation pathway. In these frameworks, data centers can coordinate with nearby renewable generators, storage systems, microgrids, and other flexible consumers through localized prices or automated energy transactions. A data center may increase consumption when local renewable production is abundant, reduce demand during distribution-network congestion, or exchange flexibility with other local participants. Compared with centralized wholesale-market participation, these mechanisms can reflect local network conditions more directly. Nevertheless, their practical implementation requires reliable communication, interoperable control platforms, transparent settlement rules, cybersecurity protection, and clear allocation of network costs.

The suitability of a data center for different market products depends on the response characteristics of its internal resources. Workload and cooling flexibility are generally more suitable for energy shifting, local flexibility, and slower demand-response products, whereas UPS batteries and BESSs are better suited to fast ancillary services because of their rapid and controllable power response. Capacity-market participation requires longer and more reliable availability, while transactive and local-market participation depends strongly on regional market design and distribution-level coordination. Therefore, technical flexibility alone does not guarantee market participation. The available response duration, baseline accuracy, metering, aggregation, battery degradation, SLA constraints, and preservation of emergency reserve must all be considered when determining the economically and operationally feasible market role of a data center [35,42,77,79–87].

Although the technical literature demonstrates that data centers can participate in demand response, reserve, and ancillary-service markets, most published results remain based on optimization models or simulation studies. Actual participation is more restrictive because qualification rules may require minimum bid sizes, telemetry, response accuracy, sustained delivery, baseline verification, and penalties for non-performance. In addition, the most profitable service is not necessarily the most appropriate service for a mission-critical facility. Fast regulation may create attractive revenue but also increase battery cycling, while energy-market arbitrage may conflict with backup-reserve preservation. Therefore, market participation should be evaluated jointly from the perspectives of technical qualification, reliability risk, degradation cost, and revenue uncertainty.

4.3. Technical Grid-Support Services and Feasibility

Large-scale data centers may provide different grid-support services through controllable workloads, server power capping, cooling adjustment, UPS systems, dedicated battery energy storage systems, backup generation, and converter-interfaced power equipment. However, these services differ substantially in activation time, delivery duration, control variable, and technical qualification. Workload scheduling is generally more suitable for slower demand response and congestion management, whereas UPS systems and BESSs are more appropriate for fast active-power regulation. Voltage support and inertia-related services additionally require sufficient converter capacity and dedicated control functions. Therefore, data-center flexibility should not be treated as a single interchangeable product across all grid services [8,35,42,77,79–83].

Frequency regulation requires repeated upward and downward active-power adjustments to correct short-term power imbalance. Data centers can support this service by modulating UPS or BESS power, applying server power capping, or adjusting a limited proportion of flexible demand. UPS systems and batteries are particularly suitable because converter-based storage can provide fast and accurate bidirectional active-power control. Earlier studies have also demonstrated the technical potential of data centers to provide ancillary and regulation services through server power control and coordinated demand adjustment [77,80–83]. Nevertheless, sustained regulation increases battery cycling and may reduce the reserve margin retained for emergency operation. Therefore, the regulation capacity offered to the grid should be coordinated with state-of-charge limits, degradation cost, and backup-duration requirements.

Fast frequency response operates over a shorter time scale and is intended to arrest rapid frequency decline immediately after a disturbance. Converter-interfaced UPS systems and BESSs can rapidly reduce grid import or inject active power when sufficient power headroom and stored energy are available. Synthetic inertia is related to fast frequency response but involves a more specific control mechanism. In this case, the converter changes active-power output according to measured frequency deviation or rate of change in frequency. Such a function cannot be provided through conventional workload shifting alone. It requires high-speed frequency measurement, dedicated converter control, sufficient energy headroom, and coordination with protection and interconnection requirements. Consequently, synthetic-inertia capability should be attributed only to data centers equipped with suitably controlled converter-based resources rather than to data-center loads in general [27,28,66].

Voltage support is mainly provided through reactive-power control or coordinated active- and reactive-power regulation at the point of interconnection. Grid-connected UPS and BESS converters may provide reactive power when their apparent-power capacity is not fully occupied by active-power transfer. However, the available capability is limited by converter rating, internal power architecture, local voltage sensitivity, and utility interconnection requirements. Voltage support is also strongly location-dependent. A response that is valuable on one constrained feeder may have little effect elsewhere in the network. In addition, recent large-load event analyses indicate that data-center loads may themselves be sensitive to voltage disturbances, which further emphasizes the need to coordinate converter support, protection settings, and load-transfer behavior [17,66,88].

Data centers may also participate in spinning and non-spinning reserve. Spinning reserve generally requires a resource to remain online and capable of increasing generation or reducing demand within a specified response period. UPS systems, BESSs, and rapidly controllable loads may provide an equivalent response by increasing battery discharge or reducing grid demand. Non-spinning reserve can be supplied by resources that are not initially delivering power but can become available after activation, including standby storage, backup generators, and prequalified curtailable workloads. However, generator startup time, fuel availability, emissions, maintenance requirements, and minimum operating duration may restrict frequent deployment. In both cases, the reserve offered to the power system must remain compatible with the energy retained for mission-critical backup. Organized markets already distinguish among regulation, synchronized reserve, and non-synchronized reserve, and storage resources may participate only when they satisfy the applicable qualification and performance requirements [42,77,83–87].

Congestion management is mainly a locational and scheduling service rather than a purely fast-response service. Data centers can contribute by shifting delay-tolerant workloads away from constrained periods, migrating workloads among regions, charging storage during unconstrained hours, and reducing demand at heavily loaded connection points. The effectiveness of these measures depends on locational grid signals, communication capacity, spare computing resources, network-aware scheduling, and the correlation of renewable availability across different sites. Poorly coordinated migration may simply transfer congestion from one location to another. Recent work on coordinated traffic–data–power scheduling therefore indicates that workload flexibility should be evaluated together with power-network constraints rather than only from the perspective of electricity price or carbon intensity [38–41,89].

Other emerging services include ramp-rate control, peak-capacity support, islanded microgrid operation, restoration support, and, in specially designed facilities, black-start assistance. Ramp smoothing and peak support can be achieved through coordinated storage, cooling, and workload control. Islanded operation requires additional protection, local generation, energy storage, and grid-forming or equivalent voltage–frequency control capability. Black-start and restoration support impose even stricter technical requirements and cannot be assumed for conventional data centers without dedicated equipment, operating procedures, and validation tests. Overall, the service portfolio of a data center should be determined by resource type, converter capability, available power and energy headroom, response speed, delivery duration, local network conditions, market qualification rules, and reliability obligations [42,66,67].

The above comparison shows that different grid services require different technical resources and control capabilities. Fast frequency-related services are mainly supported by UPS systems and BESSs, whereas workload scheduling and geographical migration are more suitable for slower demand response and congestion management. Voltage support, synthetic inertia, islanded operation, and restoration services require dedicated converter, protection, and microgrid functions and therefore cannot be generalized to all data centers. Technical capability and operational availability, rather than installed power capacity alone, should be used to determine whether a data center can reliably provide a specific grid service.

4.4. Practical Implementations and Industrial Case Studies

Although a large proportion of the literature on grid-interactive data centers is based on optimization and simulation, several industrial projects have demonstrated that selected forms of flexibility can be implemented in practice. These projects mainly involve grid-interactive UPS systems, carbon-aware workload scheduling, and low-carbon backup-power technologies. They provide useful evidence of technical feasibility, while also showing that operational implementation is strongly dependent on facility architecture, market rules, reliability requirements, and cooperation among data-center operators, utilities, and technology providers.

One representative example is Microsoft's data-center deployment in Dublin, Ireland. The facility uses grid-interactive UPS technology that allows batteries originally installed for emergency backup to exchange power with the electricity system in real time. In cooperation with the grid-service provider Enel X, the UPS batteries can respond to variations in system conditions and support grid-frequency stability, particularly when wind and solar generation fluctuate. This project demonstrates that backup batteries can provide grid services without abandoning their primary reliability function. However, practical operation still requires sufficient emergency reserve, appropriate battery and converter controls, market qualification, and coordination between grid-service commitments and data-center continuity requirements [42,82].

A second practical pathway is carbon-aware computing. Google has implemented scheduling approaches that shift selected computing activities toward times and locations with cleaner electricity. Carbon-intensity forecasts are used to guide the execution or placement of flexible computing tasks, while operational constraints and service requirements are retained. This case shows that workload flexibility can be implemented through software scheduling rather than through direct electrical power injection. Nevertheless, only workloads with sufficient temporal or spatial flexibility can be shifted, and the achievable benefit depends on the availability of reliable carbon-intensity data, spare computing capacity, communication conditions, and the differences in electricity mixes among regions [48].

A third industrial example concerns the replacement or partial substitution of diesel backup generation. Microsoft reported a demonstration in which hydrogen fuel cells supplied a row of data-center servers continuously for 48 h. This project illustrates the potential of fuel cells to provide longer-duration and lower-direct-emission backup power for mission-critical facilities. However, the demonstration does not by itself establish commercial readiness for all large-scale data centers. Fuel availability, hydrogen production pathways, storage and safety requirements, capital cost, system efficiency, maintenance, and large-scale integration must still be evaluated before widespread deployment [67].

These cases demonstrate that workload scheduling, UPS-based grid interaction, and alternative backup-power technologies have moved beyond purely theoretical analysis. However, the level of evidence remains uneven. Carbon-aware scheduling has been implemented in large cloud-computing systems, while grid-interactive UPS systems have reached operational deployment in selected markets. By contrast, some low-carbon backup technologies remain at the demonstration stage. Therefore, industrial case studies should be classified according to their maturity, including operational deployment, pilot project, and technology demonstration. The available evidence also confirms that flexibility cannot be transferred directly from one data center to another without considering differences in

https://doi.org/10.3390/en19153555

market design, grid conditions, workload composition, internal power architecture, and reliability requirements.

5. Operational Impacts on Power Systems and Data Centers

The grid interaction of large-scale data centers creates not only new flexibility opportunities, but also a series of operational effects on both power systems and the data centers themselves. In renewable-rich power systems, these effects appear in several connected aspects. These aspects include net-load shaping, reserve demand, local network stress, reliability performance, service quality, operating cost, and carbon-related results. At the same time, the actual effect of data-center flexibility depends on two key factors. One factor is how well internal resources are coordinated. The other factor is whether grid-support actions can be carried out without violating strict requirements on power continuity and computing performance. Therefore, this section reviews the main operational impacts of large-scale data-center participation. The discussion focuses on three aspects: power-system operation, data-center reliability and service quality, and broader economic and environmental performance.

5.1. Impacts on Grid Operation and Stability

Large-scale data centers affect grid operation first because of their huge electricity demand and highly concentrated connection pattern. When new facilities are connected as loads in the range of tens of megawatts, hundreds of megawatts, or even gigawatts, they can change local demand profiles, raise feeder and transformer loading, and influence regional adequacy and dispatch patterns. Early studies on cloud–grid coordination and geographical load balancing had already shown that data centers are not neutral additions to the grid, as shown in Figure 7. Depending on how operators arrange computing demand across time and space, data centers can either increase peak demand and operating stress or help reduce them [45,46,54].

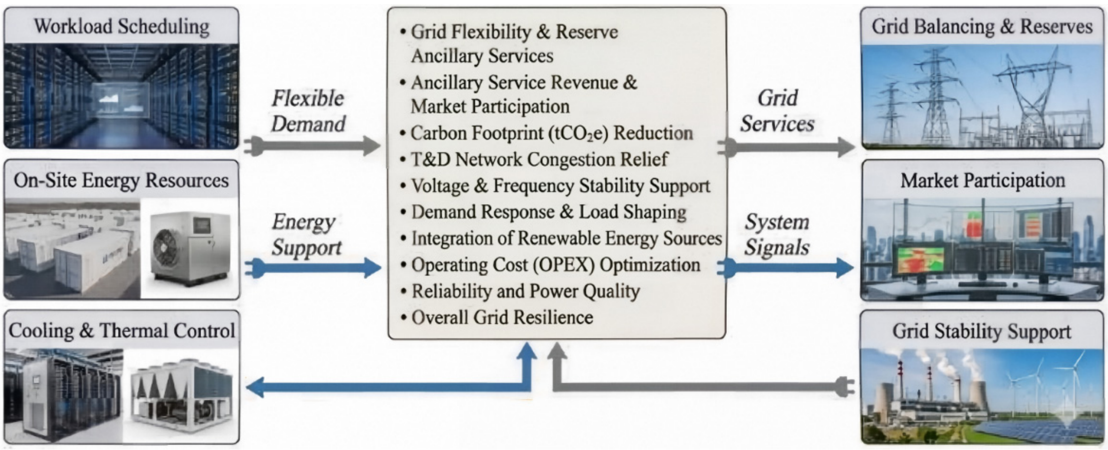

Figure 7. Impacts on grid operation and stability.

From the operational point of view, data centers can have both positive and negative effects on grid performance. If operators coordinate workload migration, demand response, and internal energy resources well, these facilities can reduce coincident peaks, smooth part of the net-load curve, and improve the time match between electricity demand and variable renewable generation. This effect can reduce short-term balancing pressure and lower the need for more expensive flexibility from other resources. However, these benefits are not automatic. The same large and rapidly changing loads can also worsen local congestion, make ramps steeper, and shift operating stress from one network area to another if flexibility actions are poorly timed or not well coordinated [8,35,42,90,91].

The impact on stability is especially important. Recent reliability studies identify data centers and other emerging large loads as a major source of rapid load growth and operating uncertainty. More importantly, event analysis shows that some data-center loads are highly sensitive to voltage changes. During transmission faults that are cleared under normal conditions, groups of facilities may switch to backup systems almost at the same time. This behavior can cause sudden and unexpected drops in bulk-system demand. Such a response creates new risks for planning and operation because traditional load models may not fully describe these characteristics, especially when several facilities are located in the same area and face the same disturbance [9,17,88].

These findings show that the effect of data centers on grid operation and stability is becoming increasingly two-sided. On the one hand, data centers can support the grid through peak reduction, renewable-aware scheduling, and fast internal electrical control. On the other hand, they can introduce new operating volatility if their load behavior is not modeled accurately or if protective actions trigger synchronized responses. For this reason, recent studies pay more attention to mitigation through better multi-network scheduling and stronger electrical interfaces. Researchers have proposed coordinated workload-shifting methods that directly consider power-network effects to reduce the grid impact of proactive data-center control. Newer studies on grid-forming battery energy storage systems used as line-interactive UPS resources also suggest a possible way to improve local electrical support during disturbances [66,89].

5.2. Impacts on Reliability, Resilience, and Service Quality

As shown in Table 9, the flexibility of large-scale data centers must always be judged against their main task, which is to provide continuous digital services with very high availability. For this reason, the impact of grid interaction cannot be evaluated only from the perspective of the power system. Researchers also need to examine its effect on the data center itself, especially in terms of reliability, resilience, and service quality. In practice, these three aspects are closely connected, but they are not exactly the same. Reliability mainly refers to the ability to maintain stable power supply, cooling, and IT operation under normal disturbances. Resilience focuses more on the ability to withstand wider or longer disruptions, adapt during the event, and recover afterward. Service quality is related to whether the facility can still meet SLAs, QoS targets, and user-performance expectations while flexibility measures are being used. Existing standards and guidelines show that data-center design is strongly shaped by redundancy, maintainability, fault tolerance, thermal compatibility, and different availability targets. This means that any attempt to extract flexibility must always remain secondary to mission-critical continuity needs [6,49,92].

From this point of view, flexible operation can bring both benefits and risks. On the positive side, coordinated demand response, workload management, and energy-aware service policies can improve operating adaptability. These methods can help operators respond to power-system stress without directly interrupting critical services. Several review studies and service-oriented analyses show that this is possible when operators only adjust delay-tolerant tasks and when contracts or operating rules clearly define the acceptable range of adjustment. However, these studies also show that flexibility always comes with a cost. If operators apply workload shifting, curtailment, or energy-aware scheduling too aggressively, they may increase the chance of SLA violations, reduce QoS, or create financial and reputational losses. Therefore, the effect of flexibility on reliability depends strongly on workload type, service importance, and control accuracy, rather than on electrical size alone [8,35,52,55].

Table 9. Summary of reliability, resilience, and service-quality impacts of grid-interactive data centers.

Aspects	Key Points	References
Reliability basis	• High-availability target • Redundancy dependence • Fault-tolerance requirement • Thermal compatibility	[6,49,92]
Service-quality impacts	• SLA compliance • QoS preservation • Penalty risk • Performance degradation	[8,35,52,55]
Resilience effects	• Outage ride-through • Islanded operation • Backup preservation • Recovery capability	[64,67,70,93]
Cross-layer influences	• Power continuity • Thermal interaction • Network reliability • Control coordination	[60,94]

The impact on resilience is also mixed. On the one hand, on-site generation, UPS systems, BESSs, and microgrid-capable structures can improve the ability of a data center to ride through outages, separate from the grid, and continue serving critical loads during wider disturbances. This is one reason why many studies view data centers as good candidates for advanced microgrid deployment. On the other hand, when operators use backup-oriented resources for market participation or grid services, the reserve margin available for emergency operation may be reduced. Recent studies on storage-based flexibility show that battery backup duration, internal power-supply structure, and dynamic UPS configuration all affect the real reliability outcome. In other words, the resilience value of flexible operation does not depend only on whether extra resources exist. It also depends on whether these resources are kept available, set correctly, and able to recover when severe events happen [64,67,70,93].

Service quality is affected not only by electrical continuity, but also by network performance, thermal conditions, and coordinated infrastructure control. Large-scale operating evidence shows that network reliability is itself an important part of end-service continuity in modern data centers. At the same time, thermal-aware studies show that energy-saving or flexibility-oriented workload control can create unexpected reliability problems if operators push server consolidation or reduce cooling without considering related thermal effects. This means that the service-quality impact of flexibility naturally crosses several layers. Electrical, thermal, computing, and network subsystems all shape the final user experience. Overall, the impact of grid-interactive operation on reliability, resilience, and service quality is conditional rather than fixed. Well-coordinated flexibility can improve adaptability and support resilient operation. Poorly coordinated flexibility, however, can weaken backup readiness, narrow thermal safety margins, and raise the risk of service degradation [60,94].

5.3. Economic and Environmental Impacts

The sustainability performance of grid-interactive data centers should be evaluated using multiple indicators. Common indicators include total carbon emissions, carbon intensity, avoided emissions, and renewable-energy utilization. Total carbon emissions represent the absolute greenhouse-gas emissions associated with electricity consumption and on-site energy use, whereas carbon intensity expresses emissions per unit of electricity consumption or computing service. These indicators are complementary because carbon intensity may decrease while total emissions continue to rise as computing demand increases [95,96].

Carbon-aware operation should also distinguish average carbon intensity from marginal carbon intensity. Average carbon intensity reflects the average generation mix and is commonly used for reporting, whereas marginal carbon intensity is more suitable for evaluating the emission effects of short-term workload shifting, battery charging, and demand-response actions. In addition, sustainability assessment should clearly define the system boundary. Operational emissions mainly include grid electricity and on-site fuel use, while lifecycle emissions also include the manufacture, replacement, and disposal of servers, batteries, cooling systems, and power equipment [96]. Therefore, future studies should report the carbon-factor source, time resolution, functional unit, and assessment boundary.

The economic and environmental effects of flexibility in large-scale data centers are closely connected. From the economic perspective, flexible operation can reduce electricity costs by moving demand away from high-price periods, cutting coincident peaks, and making better use of on-site resources. From the environmental perspective, the same flexibility can shift computing demand to periods or locations with lower carbon intensity and greater renewable-energy availability. For this reason, recent review studies point out that the value of data-center flexibility should not be judged only by its support to the grid. Researchers should also consider its role in improving cost efficiency, renewable-energy use, and emissions performance at both the facility level and the power-system level [8].

The economic value of data-center flexibility can arise from several different revenue and cost-saving pathways. These include reducing energy purchases during high-price periods, lowering demand charges and coincident peak charges, participating in demand-response and ancillary-service markets, avoiding renewable-energy curtailment, and delaying local grid or facility-capacity expansion. Workload scheduling and cooling control mainly create value by changing the timing of electricity consumption, whereas UPS systems and BESSs can additionally provide fast-response services and energy arbitrage. Backup generation may reduce outage-related losses or provide capacity support, but its routine economic use is generally restricted by fuel, maintenance, and emission costs. Therefore, the economic value of each resource depends on the specific tariff structure, market product, operating time scale, and technical qualification requirements.

Economic assessment should also include the costs created by flexibility provision. For workload shifting, these costs may include delayed task completion, migration overhead, additional communication demand, accelerator underutilization, and possible SLA penalties. Cooling flexibility may create thermal rebound, additional equipment wear, or reduced operating margins. Battery-based services introduce degradation, replacement cost, conversion losses, and the opportunity cost of preserving emergency backup capacity. Backup generators involve fuel consumption, startup cost, maintenance, emission compliance, and minimum operating-time constraints. Consequently, gross electricity savings or market revenues do not directly represent net economic benefit. A complete assessment should consider both direct operating costs and indirect reliability and service-related costs [13,51,64,73,74].

From a comparative perspective, workload and cooling flexibility generally have lower direct capital requirements because they make use of existing computing and thermal infrastructure, but their economic value is strongly constrained by service and thermal requirements. UPS and BESS resources provide more controllable and market-compatible flexibility, but their profitability depends on battery degradation, reserve preservation, electricity-price spreads, and ancillary-service payments. Backup generation offers high resilience value but is usually less attractive for frequent market operation because of fuel and environmental costs. The large differences in economic results reported in the literature are therefore mainly caused by different assumptions regarding tariffs, market prices, workload flexibility, battery lifetime, outage probability, carbon cost, and system boundaries. Future techno-economic studies should report these assumptions transparently and distinguish gross revenue, operating cost savings, lifecycle cost, and net economic benefit.

Early operational studies had already shown that the combination of workload shifting and local energy resources can produce clear economic and environmental gains. One widely cited study reported that coordinated workload shifting together with local generation reduced energy cost by about 35–40% and reduced emissions by about 10–15% under the studied conditions [54]. Related studies on renewable-powered data centers also showed that intelligent management of workloads and energy sources can lower both carbon emissions and electricity-related costs. However, the size of this benefit depends strongly on renewable intermittency and on how well different resources are coordinated [53]. These findings suggest that economic and environmental improvements are often optimized together rather than treated as fully separate goals.

More recent studies have extended this idea beyond simple electricity-cost reduction and moved toward carbon-aware operation. Carbon-aware scheduling frameworks show that operators can delay or relocate temporally flexible computing tasks to periods with lower grid carbon intensity. In this way, they can reduce electricity-related carbon emissions and, in some cases, also reduce power-infrastructure costs [48]. Under carbon-pricing conditions, the value of flexible data-center demand becomes even more evident. Studies on Internet data centers (IDCs) participating in electricity-network transition planning under carbon-pricing policies show that temporal and spatial demand adjustment can help relieve capacity pressure, support renewable-energy integration, and improve the broader economic efficiency of low-carbon network development [97]. In this sense, data-center flexibility can create value not only for operators, but also for the wider power system.

At the same time, these benefits are not guaranteed. Market prices, carbon signals, battery aging, communication burden, and service constraints all influence whether flexibility improves the final outcome. Recent coordination studies show that joint optimization between data centers and external aggregators can reduce data-center energy cost and also lower wider social cost. This result suggests that well-designed coordination frameworks can create shared economic benefits beyond the boundary of the facility itself [98]. However, recent review studies also warn that flexibility actions may only transfer costs or emissions instead of truly reducing them if they are not aligned with real renewable conditions, local grid limits, or SLAs [35]. Overall, the economic and environmental effects of data-center flexibility can be significant, but their actual value depends on careful coordination among workload scheduling, on-site resource management, and external electricity and carbon signals [8,35,48,53,54,97,98].

The economic and environmental conclusions in the existing literature are also sensitive to the selected system boundary and evaluation method. A reduction in facility-level electricity cost does not necessarily imply a reduction in total system cost, and temporal workload shifting may relocate rather than eliminate emissions if the destination region has a carbon-intensive generation mix. Similarly, operating batteries or backup generators may reduce grid-import peaks while introducing lifecycle degradation or local emissions. Therefore, both marginal grid carbon intensity and full lifecycle impacts should be considered when assessing low-carbon benefits. Results obtained from average electricity prices or average carbon factors may not accurately represent the real system-level value of flexible operation.

Overall, the economic feasibility of grid-interactive data centers should be evaluated through lifecycle and risk-aware analysis rather than through electricity-cost reduction alone. The most suitable flexibility portfolio is the one that maximizes net value after accounting for service penalties, degradation, backup preservation, operating costs, and market uncertainty.

5.4. Risks, Rebound Effects, and Cross-System Trade-Offs

The flexibility of data centers does not automatically produce system-wide benefits. Geographical workload migration may transfer electricity demand toward regions with lower prices or carbon intensity but limited transmission, substation, or feeder capacity. Moreover, multiple data centers responding simultaneously to the same price, frequency, or carbon signal may create synchronized load ramps and new congestion. Therefore, local network constraints, resource diversity, and coordinated activation and recovery should be considered in flexibility scheduling [46,54,89].

Rebound effects are another important risk. Deferred computing workloads must eventually be completed, while temporary cooling-load reductions may be followed by higher compressor, pump, or fan demand when thermal conditions are restored. Consequently, an apparent demand reduction during the response period may only shift the peak to a later time. Flexibility assessment should therefore consider the complete activation and recovery process, including energy payback, secondary peaks, and changes in workload-completion time [47,49,50,61,62].

Electrical flexibility may also create hidden reliability and environmental costs. Frequent cycling of UPS batteries can accelerate degradation and reduce emergency-backup capability unless state of charge, state of health, reserve margins, and replacement costs are explicitly considered [54,64,70,71]. Similarly, routine operation of diesel or gas backup generators may increase local air pollution and greenhouse-gas emissions, even when grid electricity consumption is reduced. Grid benefits should therefore be evaluated together with battery degradation, reserve erosion, fuel consumption, start-up losses, maintenance requirements, and lifecycle emissions [68,69].

Finally, the operational objectives of data centers and power systems may not always be aligned. Data-center operators prioritize service continuity, latency, throughput, SLA compliance, and commercial risk, whereas system operators focus on reliability, congestion management, reserve adequacy, and system-wide cost. Price-based incentives alone may therefore be insufficient to guarantee dependable grid support. Credible participation requires clearly defined contractual limits, performance baselines, opt-out conditions, recovery rules, and verification mechanisms [6,8,35,49,52,55,92].

6. Discussion About Challenges and Future Research Directions

Although large-scale data centers have shown increasing potential as flexible and grid-interactive resources, many important challenges still remain before this potential can be fully and reliably used in power systems with high renewable penetration. On the one hand, data centers contain several forms of flexibility. These include workload scheduling, cooling adjustment, UPS support, BESS operation, and backup power coordination. On the other hand, these resources differ greatly in response speed, support duration, controllability, and operating purpose. This difference makes integrated coordination difficult. In addition, the real value of flexibility is always limited by strict requirements on reliability, SLAs, QoS, thermal safety, and emergency readiness. Therefore, the central issue is not simply whether data centers can support the grid. The more important question is how this support can be delivered under realistic operating limits.

Another major challenge comes from the stronger coupling between renewable-rich power systems and large-scale digital infrastructure. As renewable penetration continues to rise, power systems require more flexibility that is fast, multi-timescale, and able to deal with uncertainty. At the same time, data-center loads are becoming larger, more dynamic, and more geographically concentrated, especially with the rapid growth of AI and HPC workloads. This trend creates new tension between grid-support goals and computing-service responsibilities. It also reveals the limits of many current modeling methods. Existing studies often treat data centers either as simplified large loads or as separate flexible resources. These approaches often fail to fully describe the internal coupling among electrical, thermal, and computing processes, as well as the interaction between data centers and external markets.

Future research should first develop more accurate and scalable models that jointly represent electrical, thermal, computing, communication, and reliability processes in large-scale data centers. Existing studies often simplify one or more of these domains, which can lead to overestimated flexibility and underestimated operational risk. Future models should explicitly represent workload heterogeneity, cooling dynamics, battery degradation, backup requirements, network constraints, and voltage-sensitive load behavior. In addition, standardized benchmark systems, workload datasets, thermal models, and evaluation scenarios are needed to enable fair comparison among different scheduling and control methods. Without common benchmarks, the flexibility values reported by different studies remain difficult to compare or reproduce.

A second research direction is the development of multi-timescale coordination methods under uncertainty. Data-center resources operate over substantially different time scales: UPS systems and BESSs can respond within milliseconds or seconds, cooling systems act over minutes to hours, and workload scheduling may extend from hours to days. Future energy management systems should coordinate these resources through day-ahead scheduling, rolling optimization, and real-time control. Renewable-generation uncertainty, workload arrivals, electricity prices, network conditions, and service requests should be considered jointly rather than independently. Hybrid approaches that combine stochastic or robust scheduling with model predictive control may provide a practical balance between economic performance, operational feasibility, and reliability.

The rapid growth of AI-driven data centers creates a need for workload-specific flexibility assessment. Future studies should distinguish large-scale training, fine-tuning, batch inference, and real-time inference because these workloads have different delay tolerance, migration capability, power variation, and cooling requirements. In particular, the flexibility obtained through checkpointing, training interruption, accelerator power capping, request routing, and model placement should be quantified together with restart cost, synchronization overhead, data locality, and accelerator utilization. Future research should also investigate whether coordinated scheduling among geographically distributed AI facilities can reduce renewable curtailment and congestion without transferring network stress or violating latency and service-quality requirements.

Another important direction is the development of advanced electrical interfaces for grid support. Most current studies focus on active-power scheduling, while the potential roles of grid-interactive UPS systems, grid-forming BESSs, reactive-power control, fast frequency response, synthetic inertia, and islanded operation remain insufficiently validated. Future research should examine converter control, protection coordination, fault ride-through, synchronization, and transition between grid-connected and islanded modes. Particular attention should be given to simultaneous voltage-sensitive load reductions, because correlated protective actions by multiple data centers may create unexpected bulk-system disturbances. These functions require hardware-in-the-loop testing, pilot-scale validation, and coordination with utility interconnection requirements.

Future market and regulatory frameworks should recognize that data-center flexibility is conditional rather than fully dispatchable. Qualification rules should account for backup-reserve preservation, battery degradation, SLA and QoS constraints, response duration, locational value, and non-performance risk. New mechanisms are also needed for local flexibility markets, congestion management, capacity support, and coordinated participation through aggregators. At the same time, future studies should move beyond simulation-based economic analysis and provide pilot projects, field measurements, and long-term operational evidence. Economic and environmental assessment should include electricity cost, market revenue, battery replacement cost, backup-fuel use, marginal carbon intensity, lifecycle emissions, water consumption, and possible rebound effects. This broader evaluation is necessary to determine whether flexibility actions create genuine system-level benefits or merely shift costs and emissions across time and location.

Overall, future research should move from isolated resource optimization and idealized simulation toward integrated, uncertainty-aware, and experimentally validated coordination. The key objective is not to maximize the theoretical flexibility of data centers, but to identify the flexibility that can be delivered reliably, economically, and sustainably without weakening mission-critical services. Progress in standardized modeling, multi-timescale control, advanced converter interfaces, market design, and field validation will determine whether large-scale data centers can become dependable grid-supportive resources in high-renewable-penetration power systems.

7. Conclusions

This paper reviewed large-scale data centers in power systems with high renewable penetration from the perspectives of flexible resources, grid interaction, and operational impact. The review shows that large-scale data centers should not be regarded only as rapidly growing electricity loads. With workload scheduling, cooling adjustment, UPS systems, BESSs, and backup power resources, they are gradually becoming grid-interactive facilities with multi-layer flexibility. At the same time, this review also shows that the role of data centers has a clear two-sided nature. When operators coordinate flexibility properly, data centers can help reduce peaks, improve renewable-energy use, support short-term balancing, and strengthen resilient operation. However, these same facilities can also bring new challenges to grid operation, system stability, and power-system planning. The main reasons are their large electrical scale, concentrated grid connection, and strict requirements on reliability, SLAs, and QoS.

Overall, the value of large-scale data centers in future power systems depends not only on how much flexibility they contain, but also on how well this flexibility is coordinated under realistic operating constraints. With continued progress in modeling, control, and market design, large-scale data centers are likely to become an important type of flexible and grid-supportive resource in renewable-rich power systems.

Author Contributions: Writing—original draft, M.C.; writing—review and editing, S.M. and H.M.; Validation and Writing—review and editing, W.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Natural Science Basic Research Program of Shaanxi (Program No. 2025JC-YBQN-711) and Aeronautical Science Foundation of China (No. 20240028053003).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Acknowledgments: The authors acknowledge the use of ChatGPT 4.5 (OpenAI) solely for language polishing and figure improvements. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. International Energy Agency. *Energy and AI*; IEA: Paris, France, 2025.
2. International Energy Agency. *The Power of Transformation: Wind, Sun and the Economics of Flexible Power Systems*; OECD/IEA: Paris, France, 2014.
3. Rahman, M.M.; Dadon, S.H.; He, M.; Giesselmann, M.; Hasan, M.M. An Overview of Power System Flexibility: High Renewable Energy Penetration Scenarios. *Energies* **2024**, *17*, 6393. [CrossRef]
4. Makolo, P.; Zamora, R.; Lie, T.-T. The role of inertia for grid flexibility under high penetration of variable renewables—A review of challenges and solutions. *Renew. Sustain. Energy Rev.* **2021**, *147*, 111223. [CrossRef]
5. Kamiya, G.; Coroamă, V.C. *Data Centre Energy Use: Critical Review of Models and Results*; EDNA–IEA 4E TCP: Paris, France, 2025.
6. Uptime Institute. *Tier Classification System*; Uptime Institute: New York, NY, USA, 1993.
7. ASHRAE. *ASHRAE TC 9.9. Data Center Power Equipment Thermal Guidelines and Best Practices*; White Paper; ASHRAE: Peachtree Corners, GA, USA, 2016.
8. Takci, M.T.; Qadrdan, M.; Summers, J.; Gustafsson, J. Data centres as a source of flexibility for power systems. *Energy Rep.* **2025**, *13*, 3661–3671. [CrossRef]
9. North American Electric Reliability Corporation. *2024 Long-Term Reliability Assessment*; NERC: Atlanta, GA, USA, 2024.
10. Mohammadi, S.; Wang, W.; Wada, M.C.I.; Haghighi, R.; Hassan, A.; Liu, H.; Bhatnagar, A.; Chen, A.; Su, W. Grid Integration of AI Data Centers: A Critical Review of Energy Storage Solutions. *Adv. Appl. Energy* **2026**, *23*, 100291. [CrossRef]
11. Han, J.; Yan, Y.; Wang, Y.; Han, K.; Han, Y.; Lin, J. Dual-time scale collaborative optimization of data center integrated energy systems. *J. Energy Storage* **2025**, *119*, 116244. [CrossRef]
12. Fan, Y.; Zhao, J. Harnessing Flexible Spatial and Temporal Data Center Workloads for Grid Regulation Services. *arXiv* **2026**, arXiv:2602.01508.
13. Zhang, Y.; Tang, H.; Li, H.; Wang, S. Unlocking the flexibilities of data centers for smart grid services: Optimal dispatch and design of energy storage systems under progressive loading. *Energy* **2025**, *316*, 134511. [CrossRef]
14. Kumar, A.; Pindoriya, N.M. Toward Sustainable Data Center Operation: A Review on Renewable Integration, Energy Flexibility, and Intelligent Management. *Renew. Sustain. Energy Rev.* **2026**, *230*, 116664. [CrossRef]
15. Zhang, Y.; Tang, H.; Li, H.; Wang, S. Integration and Interaction of Next-Generation AI-Focused GPU Data Centers with Smart Grids and District Energy Systems: A Review. *Renew. Sustain. Energy Rev.* **2025**, *224*, 116097. [CrossRef]
16. Huang, P.; Copertaro, B.; Zhang, X.; Shen, J.; Löfgren, I.; Rönnelid, M.; Fahlen, J.; Andersson, D.; Svanfeldt, M. A Review of Data Centers as Prosumers in District Energy Systems: Renewable Integration and Waste Heat Utilization. *Appl. Energy* **2020**, *258*, 114109. [CrossRef]
17. North American Electric Reliability Corporation. *Characteristics and Risks of Emerging Large Loads*; NERC: Atlanta, GA, USA, 2025.
18. Uptime Institute. *Global Data Center Survey 2024*; Uptime Institute: New York, NY, USA, 2024.
19. American Society of Heating; Refrigerating and Air-Conditioning Engineers. *Emergence and Expansion of Liquid Cooling in Mainstream Data Centers*; ASHRAE: Peachtree Corners, GA, USA, 2021.
20. International Energy Agency (IEA). *Integrating Solar and Wind*; IEA: Paris, France, 2024.
21. Ela, E.; Milligan, M.; Kirby, B. *Operating Reserves and Variable Generation: A Comprehensive Review of Current Strategies, Studies, and Fundamental Research on the Impact That Increased Penetration of Variable Renewable Generation Has on Power System Operating Reserves*; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2011.
22. Bird, L.; Milligan, M.; Lew, D. Integrating Variable Renewable Energy: Challenges and Solutions. *Environ. Res. Lett.* **2013**, *8*, 034028.
23. Bessa, R.J.; Moreira, C.; Silva, B.; Matos, M.A. Handling Renewable Energy Variability and Uncertainty in Power Systems Operation. *Wiley Interdiscip. Rev. Energy Environ.* **2014**, *3*, 156–178.
24. Ahmed, A.; Khalid, M. A Review on the Selected Applications of Forecasting Models in Renewable Power Systems. *Renew. Sustain. Energy Rev.* **2019**, *100*, 9–21. [CrossRef]
25. Huang, B.; Krishnan, V.; Hodge, B. Analyzing the Impacts of Variable Renewable Resources on California Net-Load Ramp Events. In *NREL Conference Paper*; NREL: Golden, CO, USA, 2018.

26. Milligan, M.; Cochran, J.; Denholm, P. *Advancing System Flexibility for High Penetration Renewable Integration*; NREL: Golden, CO, USA, 2015.
27. Denholm, P.; Sun, Y.; Mai, T. *Inertia and the Power Grid: A Guide Without the Spin*; NREL: Golden, CO, USA, 2020.
28. Dörfler, F.; Groß, D. Control of Low-Inertia Power Systems. *Annu. Rev. Control Robot. Auton. Syst.* **2023**, *6*, 415–445. [CrossRef]
29. Song, J.; Zhou, X.; Zhou, Z.; Wang, Y.; Wang, Y.; Wang, X. Review of Low Inertia in Power Systems Caused by High Proportion of Renewable Energy Grid Integration. *Energies* **2023**, *16*, 6042. [CrossRef]
30. Mohandes, B.; El Moursi, M.S.; Hatziargyriou, N.; El Khatib, S. A Review of Power System Flexibility with High Penetration of Renewables. *IEEE Trans. Power Syst.* **2019**, *34*, 3140–3155. [CrossRef]
31. Gong, C.; Hu, Y.; Gao, J.; Wang, Y.; Yan, L. An Improved Delay-Suppressed Sliding-Mode Observer for Sensorless Vector-Controlled PMSM. *IEEE Trans. Ind. Electron.* **2020**, *67*, 5913–5923. [CrossRef]
32. International Renewable Energy Agency (IRENA). *Power System Flexibility for the Energy Transition, Part 1: Overview for Policy Makers*; IRENA: Abu Dhabi, UAE, 2018.
33. North American Electric Reliability Corporation (NERC). *2025 Long-Term Reliability Assessment*; NERC: Atlanta, GA, USA, 2026.
34. Garg, A.; Booten, C.; VanGeet, O.; Pless, S.; Stenger, K.; Vidal, J.; Jackson, R.; Lavin, L.; Prabakar, K. *Considerations for Distributed Edge Data Centers and Use of Building Loads to Support Large Interconnections*; National Renewable Energy Laboratory (NREL): Golden, CO, USA, 2025.
35. Chen, M.; Gao, C.; Song, M.; Chen, S.; Li, D.; Liu, Q. Internet Data Centers Participating in Demand Response: A Comprehensive Review. *Renew. Sustain. Energy Rev.* **2020**, *120*, 109609.
36. Basmadjian, R.; Botero, J.F.; Giuliani, G.; Hesselbach, X.; Klingert, S.; de Meer, H. Making Data Centers Fit for Demand Response: Introducing GreenSDA and GreenSLA Contracts. *IEEE Trans. Smart Grid* **2018**, *9*, 3453–3464. [CrossRef]
37. Cao, Y.; Cheng, M.; Zhang, S.; Mao, H.; Wang, P.; Li, C.; Feng, Y.; Ding, Z. Data-Driven Flexibility Assessment for Internet Data Center towards Periodic Batch Workloads. *Appl. Energy* **2022**, *324*, 119665. [CrossRef]
38. Liu, Z.; Lin, M.; Wierman, A.; Low, S.H.; Andrew, L.L.H. Geographical Load Balancing with Renewables. In *Proceedings of the 2011 GreenMetrics Workshop*; ACM: New York, NY, USA, 2011; pp. 62–66.
39. Zheng, J.; Chien, A.A.; Suh, S. Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers. *Joule* **2020**, *4*, 2208–2222. [CrossRef]
40. Wiesner, P.; Behnke, I.; Scheinert, D.; Gontarska, K.; Thamsen, L. Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. In Proceedings of the 12th ACM International Conference on Future Energy Systems (e-Energy '21), Virtual, 28 June–2 July 2021.
41. Naoi, H.; Delage, R.; Nakata, T.; Kozai, M. Integrating Location Planning and Spatio-Temporal Workload Shifting for 100% Renewable Energy Data Centers: A Case Study in Japan. *Energy Strategy Rev.* **2025**, *61*, 101823. [CrossRef]
42. Paananen, J. Grid-Interactive Data Centers Enabling Energy Transition: Data center's hidden potential to provide essential grid services of a future power system. *IEEE Electrif. Mag.* **2023**, *11*, 26–34. [CrossRef]
43. Dayarathna, M.; Wen, Y.; Fan, R. Data Center Energy Consumption Modeling: A Survey. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 732–794. [CrossRef]
44. Ebrahimi, K.; Jones, G.F.; Fleischer, A.S. A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities. *Renew. Sustain. Energy Rev.* **2014**, *31*, 622–638. [CrossRef]
45. Mohsenian-Rad, A.-H.; Leon-Garcia, A. Coordination of Cloud Computing and Smart Power Grids. In Proceedings of the 2010 First IEEE International Conference on Smart Grid Communications, Gaithersburg, MD, USA, 4–6 October 2010; pp. 368–372. [CrossRef]
46. Liu, Z.; Lin, M.; Wierman, A.; Low, S.H.; Andrew, L.L.H. Greening Geographical Load Balancing. In Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems, San Jose, CA, USA, 7–11 June 2011; pp. 233–244. [CrossRef]
47. Liu, Z.; Chen, Y.; Bash, C.; Wierman, A.; Gmach, D.; Wang, Z.; Marwah, M.; Hyser, C. Renewable and Cooling Aware Workload Management for Sustainable Data Centers. In Proceedings of the ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems, London, UK, 11–15 June 2012; pp. 175–186. [CrossRef]
48. Radovanović, A.; Koningstein, R.; Schneider, I.; Chen, B.; Duarte, A.; Roy, B.; Xiao, D.; Haridasan, M.; Hung, P.; Care, N.; et al. Carbon-Aware Computing for Datacenters. *IEEE Trans. Power Syst.* **2023**, *38*, 1270–1280. [CrossRef]
49. ASHRAE Technical Committee 9.9. *2021 Equipment Thermal Guidelines for Data Processing Environments*, 5th ed.; ASHRAE: Peachtree Corners, GA, USA, 2021.
50. Liu, L.; Zhang, Q.; Zhai, Z.J.; Yue, C.; Ma, X. State-of-the-art on thermal energy storage technologies in data center. *Energy Build.* **2020**, *226*, 110345. [CrossRef]
51. Wang, Z.; Yin, Z.; Yang, J.; Wang, J. Coordinated optimization of distributed energy system and storage-enhanced uninterruptible power supply in data center: A three-level optimization framework with model predictive control. *Energy Convers. Manag.* **2025**, *342*, 120137. [CrossRef]

52. Klingert, S.; Schulze, T.; Bunse, C. GreenSLAs for the Energy-Efficient Management of Data Centres. In Proceedings of the 2nd International Conference on Energy-Efficient Computing and Networking (e-Energy), Oldenburg, Germany, 31 May–1 June 2011; pp. 21–30.
53. Goiri, I.; Katsak, W.; Le, K.; Nguyen, T.D.; Bianchini, R. Parasol and GreenSwitch: Managing Datacenters Powered by Renewable Energy. In Proceedings of the Eighteenth International Conference on Architectural Support for Programming Languages and Operating Systems, Houston, TX, USA, 16–20 March 2013; pp. 51–64.
54. Liu, Z.; Wierman, A.; Chen, Y.; Razon, B.; Chen, N. Data Center Demand Response: Avoiding the Coincident Peak via Workload Shifting and Local Generation. *Perform. Eval.* **2013**, *70*, 770–791. [CrossRef]
55. Amokrane, A.; Langar, R.; Zhani, M.F.; Boutaba, R.; Pujolle, G. Greenslater: On Satisfying Green SLAs in Distributed Clouds. *IEEE Trans. Netw. Serv. Manag.* **2015**, *12*, 363–376. [CrossRef]
56. Khalaj, A.H.; Halgamuge, S.K. A Review on Efficient Thermal Management of Air- and Liquid-Cooled Data Centers: From Chip to the Cooling System. *Appl. Energy* **2017**, *205*, 1165–1188. [CrossRef]
57. Zhang, Q.; Meng, Z.; Hong, X.; Zhan, Y.; Liu, J.; Dong, J.; Bai, T.; Niu, J.; Deen, M.J. A Survey on Data Center Cooling Systems: Technology, Power Consumption Modeling and Control Strategy Optimization. *J. Syst. Archit.* **2021**, *119*, 102253. [CrossRef]
58. Wan, J.; Gui, X.; Kasahara, S.; Li, J. Air Flow Measurement and Management for Improving Cooling and Energy Efficiency in Raised-Floor Data Centers: A Survey. *IEEE Access* **2018**, *6*, 48884–48898. [CrossRef]
59. Vasic, N.; Scherer, T.; Schott, W. Thermal-Aware Workload Scheduling for Energy Efficient Data Centers. In Proceedings of the 7th International Conference on Autonomic Computing (ICAC), Washington, DC, USA, 7–11 June 2010; pp. 169–174.
60. MirhoseiniNejad, S.M.; Moazamigoodarzi, H.; Badawy, G.; Down, D.G. Joint Data Center Cooling and Workload Management: A Thermal-Aware Approach. *Future Gener. Comput. Syst.* **2020**, *104*, 174–189. [CrossRef]
61. Cai, S.; Gou, Z. Towards Energy-Efficient Data Centers: A Comprehensive Review of Passive and Active Cooling Strategies. *Energy Built Environ.* **2026**, *7*, 206–226. [CrossRef]
62. Alkrush, A.A.; Salem, M.S.; Abdelrehim, O.; Hegazi, A.A. Data Centers Cooling: A Critical Review of Techniques, Challenges, and Energy Saving Solutions. *Int. J. Refrig.* **2024**, *160*, 246–262. [CrossRef]
63. Narayanan, I.; Wang, D.; Mamun, A.-A.; Sivasubramaniam, A.; Fathy, H.K. Should We Dual-Purpose Energy Storage in Datacenters for Power Backup and Demand Response? In Proceedings of the HotPower 2014, Broomfield, CO, USA, 5 October 2014.
64. Yu, Y.; Shan, K.; Tang, H.; Wang, S. Reliability and economic impacts of utilizing battery energy storage in data centers for energy flexibility services in smart grids. *Energy Convers. Manag.* **2025**, *339*, 119951. [CrossRef]
65. Li, L.; Pu, Y.; Ju, Y.; Hu, Z.; Wang, Z. Coordinating Multiple UPS Batteries in Datacenter for Load Flexibility. *Energy Build.* **2026**, *359*, 117282. [CrossRef]
66. Azizi, A.; Morovati, S.; Zamani, A.; Piruzza, J.; Guo, D.; Liu, Z.; Taimela, P. Strengthening Data Center Operations Using Grid-Forming Battery Energy Storage as a Line-Interactive Uninterruptible Power Supply. *Int. J. Electr. Power Energy Syst.* **2026**, *175*, 111638. [CrossRef]
67. Sheehan, S.; Rakow, A. Evolving a Data Center Into a Microgrid: Industry perspectives and lessons learned. *IEEE Electrif. Mag.* **2023**, *11*, 16–25. [CrossRef]
68. Schweiger, J.; Macdonald, R. Techno-Economic Analysis of Decarbonized Backup Power Systems for Data Centers. *Energies* **2025**, *18*, 4388. [CrossRef]
69. Li, C.; Wang, R.; Li, T.; Qian, D.; Yuan, J. Managing Green Datacenters Powered by Hybrid Energy Supply. In Proceedings of the 11th International Conference on Autonomic Computing (ICAC), Philadelphia, PA, USA, 8–20 June 12014.
70. Ahmed, K.M.U.; Tuan, N.; Lehtonen, M.; Agbossou, A. Reliability Analysis of Internal Power Supply Architecture of Data Centers in Terms of Power Losses. *Electr. Power Syst. Res.* **2021**, *193*, 107025. [CrossRef]
71. Gong, C.; Li, Y.R.; Zargari, N.R. An Overview of Advancements in Multimotor Drives: Structural Diversity, Advanced Control, Specific Technical Challenges, and Solutions. *Proc. IEEE* **2024**, *112*, 184–209. [CrossRef]
72. Rahman, S.; Tbaileh, A.; Elwany, A. Energy Storage Systems for AI Data Centers: A Review of Benefits, Challenges, and Future Directions. *Energies* **2026**, *19*, 634.
73. Safari, A.; Blaabjerg, F.; Oshnoei, A. A research-industry perspective of battery systems technology for next-generation data centers. *J. Energy Storage* **2026**, *152*, 120386. [CrossRef]
74. Ziaei, P.; Husev, O.; Rabkowski, J. Data Centers as a Driving Force for the Renewable Energy Transition: Toward Grid-Interactive and Flexible Architectures. *Energies* **2025**, *19*, 236.
75. Urgaonkar, R.; Urgaonkar, B.; Neely, M.J.; Sivasubramaniam, A. Optimal Power Cost Management Using Stored Energy in Data Centers. In Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, San Jose, CA, USA, 7–11 June 2011; pp. 221–232.
76. Cioara, T.; Anghel, I.; Bertoncini, M.; Salomie, I.; Arnone, D.; Mammina, M.; Velivassaki, T.-H.; Antal, M. Optimized Flexibility Management Enacting Data Centres Participation in Smart Demand Response Programs. *Future Gener. Comput. Syst.* **2018**, *78*, 330–342. [CrossRef]

77. Niu, L.; Guo, Y. Enabling Reliable Data Center Demand Response via Aggregation. In Proceedings of the Seventh International Conference on Future Energy Systems (e-Energy '16), Waterloo, ON, Canada, 21–24 June 2016.
78. Sheikholeslami, S.M.; Rabiei, A.M.; Mohammad-Taheri, M.; Abouei, J. Cloud Data Center Participation in Smart Demand Response Programs for Energy Cost Minimisation. *IET Smart Grid* **2022**, *5*, 380–394. [CrossRef]
79. Wierman, A.; Liu, Z.; Liu, I.; Mohsenian-Rad, H. Opportunities and Challenges for Data Center Demand Response. In Proceedings of the International Green Computing Conference, Dallas, TX, USA, 3–5 November 2014.
80. Ghamkhari, M.; Mohsenian-Rad, H. Data Centers to Offer Ancillary Services. In Proceedings of the 2012 IEEE First International Conference on Smart Grid Communications, Tainan, Taiwan, 5–8 November 2012; pp. 436–441.
81. Aikema, D.; Simmonds, R.; Zareipour, H. Delivering ancillary services with data centres. *Sustain. Comput. Inform. Syst.* **2013**, *3*, 269–281. [CrossRef]
82. Chen, H.; Coskun, A.K.; Caramanis, M.C. Dynamic Server Power Capping for Enabling Data Center Regulation Service. In Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, USA, 2–4 October 2013.
83. Zhang, X.; Deng, N.; Ren, S.; He, Y.; Yang, Y.; Nahrstedt, K. QoS-Aware Data Center Participation in Smart Grid Regulation Service Reserve Provision. *ACM Trans. Model. Perform. Eval. Comput. Syst.* **2018**, *4*, 1–31.
84. Federal Energy Regulatory Commission. *Order No. 841: Electric Storage Participation in Markets Operated by Regional Transmission Organizations and Independent System Operators*; FERC: Washington, DC, USA, 2018.
85. PJM Interconnection. *PJM Manual 11: Energy & Ancillary Services Market Operations*; PJM Interconnection: Audubon, PA, USA, 2023.
86. Electric Reliability Council of Texas. *Load Resource Participation in the ERCOT Markets*; Official Program Page; ERCOT: Taylor, TX, USA, 2007.
87. Vertiv. *How to Maximize Revenues from Your Data Center Energy Storage with Grid-Interactive UPS*; White Paper; Vertiv: Columbus, OH, USA, 2022.
88. North American Electric Reliability Corporation. *Incident Review: Considering Simultaneous Voltage-Sensitive Load Reductions*; NERC: Atlanta, GA, USA, 2025.
89. Zhang, Y.; Ye, Y.; Zhong, H.; Wang, Z.; Tang, Z.; Kang, C. Mitigating Power Grid Impact from Proactive Data Center Workload Shifts: A Coordinated Scheduling Strategy Integrating Synergistic Traffic-Data-Power Networks. *Appl. Energy* **2025**, *377*, 124602. [CrossRef]
90. Guo, C.; Luo, F.; Cai, Z.; Dong, Z.Y. Integrated Energy Systems of Data Centers and Smart Grids: State-of-the-Art and Future Opportunities. *Appl. Energy* **2021**, *301*, 117474. [CrossRef]
91. Alaperä, I.; Honkapuro, S.; Paananen, J. Data Centers as a Source of Dynamic Flexibility in Smart Grids. *Appl. Energy* **2018**, *229*, 69–79. [CrossRef]
92. U.S. Department of Energy; Lawrence Berkeley National Laboratory. *Designing and Managing Data Centers for Resilience: Demand Response and Microgrids*; U.S. Department of Energy: Washington, DC, USA, 2019.
93. Varnavskiy, K.; Nepsha, F.; Chen, Q.; Ermakov, A.; Zhironkin, S. Reliability Assessment of the Configuration of Dynamic Uninterruptible Power Sources: A Case of Data Centers. *Energies* **2023**, *16*, 1419. [CrossRef]
94. Meza, J.; Wu, Q.; Kumar, S.; Mutlu, O. A Large Scale Study of Data Center Network Reliability. In Proceedings of the Internet Measurement Conference, Boston, MA, USA, 31 October–2 November 2018; pp. 393–407.
95. World Resources Institute; World Business Council for Sustainable Development. *GHG Protocol Scope 2 Guidance: An Amendment to the GHG Protocol Corporate Standard*; World Resources Institute: Washington, DC, USA, 2015.
96. Üçtuğ, F.G.; Ünver, T.C. Life cycle assessment-based environmental impact analysis of a tier 4 data center: A case study in Turkey. *Sustain. Energy Technol. Assess.* **2023**, *56*, 103076. [CrossRef]
97. Wan, T.; Ding, T.; Wei, Y.; Li, C.; Shahidehpour, M.; Zhu, J.; Li, F. Internet data centers participating in electricity network transition roadmap under carbon pricing policies. *Appl. Energy* **2023**, *329*, 120239.
98. Zhang, Y.; Li, J.; Zhang, C.; Liu, Z. Data center and load aggregator coordination towards power flexibility in smart grids. *Sustain. Energy Grids Netw.* **2024**, *37*, 101298.