Carbon-Aware Coordination of Power and Computing Resources in Digital Infrastructure

Ali Abubakar Muumin[1], Muhammad Waqas Latif[2], *
[1]Senior ICT officer, Zanzibar Government Printing Press Agency, Zanzibar, Tanzania
[2]Research Scholar, Tianjin Key Laboratory of Process Measurement and Control, School of Electrical and Information Engineering, Tianjin University, China
Email: lmwaqas@tju.edu.cn

Abstract-The interplay between electrical power and computing power is more important than ever in the era of artificial intelligence (AI), cloud computing, and edge computing, as data centres are becoming controllable electrical loads. The results of this survey are presented on various aspects of grid-interactive, carbon-aware coordination of electricity and computing resources over the last ten years, including multi-energy data centers, edge-cloud scheduling, computing-power networks, workload shifting, and demand response. Studies are included only if they coordinate computation with at least one electrical factor, such as price, carbon intensity, grid constraints, renewables, cooling, or storage. To differentiate market coupling, carbon-aware dispatching, networked compute orchestration, local energy assets, and metrics, a five-layer taxonomy has been proposed. For practical implementation, the key prerequisites are latency-guaranteed, nodal-carbon-aware workload-power co-dispatch, interoperable data exchange, and verifiable grid benefits.

Keywords: Electricity-computing, carbon-aware computing, computing power network, edge-cloud scheduling, grid-interactive loads.

I. Introduction

Digital infrastructure is no longer a passive electric-power recipient. These systems are now deployed at scale, impacting electrical systems, renewable curtailment, reserve procurement, and carbon emissions in large data centers, edge nodes, AI clusters, and networked computing resources. Computing workloads are also more flexible than most industrial workloads. Batch AI training jobs could be halted or moved, and inference services could be replicated across edge or cloud nodes. Cooling plants, batteries, and backups can also be scheduled based on server utilization. Therefore, the planning and operation of electricity systems and computing systems should be coordinated rather than separately optimized. Past energy-efficiency studies of data centers also reveal that total energy consumption is a function of the power used by information technology (IT) equipment and the overhead power used in energy-intensive data-center facilities [1], [2]. Recently, it has become common practice to regard a data center as a dispatchable, carbon-aware, and market-responsive load [3]–[12].

This change is spurred by three drivers. First, AI and cloud workloads represent dense and geographically concentrated electricity demand, which is straining grid-connection queues, overloading transformers, and forcing network expansion in some areas. Second, renewable production varies by location. Migrating a computing task to a region with lower carbon intensity can lead to lower emissions, but it can also increase network latency or burden the receiving electrical grid. Third, the computing-power network (CPN) paradigm considers computing, storage, and networking resources as integrated services rather than individual resources [13]–[15]. These developments have introduced the idea that electrical power

Journal of Intelligent Engineering and Informatics, Volume 1, Issue 1, Pages 70-82
Corresponding author. E-mail address: lmwaqas@tju.edu.cn
Received 17 July 2026; **revised** 23 August 2026; **accepted** 31 August 2026; **published** 11 September 2026.

systems and computing systems present shared problems in electricity-compute coordination. Current reviews are scattered across specific research areas. While data-center energy-modeling reviews focus on energy management and consumption [2], [16], [17], demand-response reviews are centered on load flexibility [13]–[15], and mobile-edge and resource-scheduling surveys are centered on offloading and edge orchestration [18]–[21]. An integrated review that takes into account electricity-market signals, workload migration targeting carbon reduction, edge-cloud orchestration, local cooling and storage capacities, and grid-level validation metrics is still limited. This gap is significant because what works for cloud scheduling is not necessarily feasible for grid operation, and what is suitable for a demand-response strategy might not meet service-level latency requirements. In line with this, this survey examines grid-interactive/computing-aware models that consider the carbon dimension and sets the scope as follows.

The studies considered should relate computational needs to price, carbon intensity, grid constraints, renewable availability, demand-response signals, cooling energy, battery operation, heat reuse, or local generation. The survey covers data-center clusters, hyperscale cloud regions, edge nodes, computing-power networks, and multi-energy data-center buildings. Pure chip-level design, thermal-only studies that do not take workload or grid interaction into account, blockchain-mining economics, and generic cloud scheduling without electricity coupling are excluded. The review approach is concept-based rather than chronological. Studies published between 2015 and 2026 are grouped according to five coordinated decisions: electricity-market response, carbon-aware workload allocation, networked computing-resource orchestration, local multi-energy flexibility, and evaluation or verification. The analysis focuses on exploitable assumptions, assumptions required during simulation, and missing interfaces that prevent coordinated operation. The proposed electricity-compute coordination loop is shown in Figure 1. The boundary of the survey is described in Table 1, and the main taxonomy is provided in Table 2. The research challenge of latency-guaranteed, nodal-carbon-aware workload-power co-dispatch is highlighted. Much of the recent literature assumes a parameterized regional-average carbon intensity and high workload mobility or treats demand response without explicitly considering power-flow constraints. For practical systems, decisions about which computing service to place on which device at what time, in the context of service latency, grid congestion, renewable variability, cooling dynamics, and auditable carbon accounting, are crucial. It is recognised, therefore, that future activities should focus on constructing interoperable testbeds, grid-connected validation, and common metering for carbon, cost, reliability, and service quality.

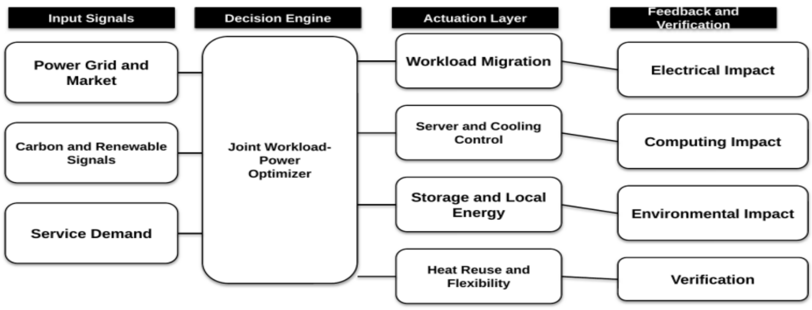

Figure 1. Proposed electricity-compute coordination framework

Figure 1 illustrates four coordinated layers: external input signals, a joint decision engine, controllable data-center actions, and feedback signals for technical verification.

Table 1. Survey boundary and inclusion logic for electricity-compute coordination.

Item	Definition used in this survey
Included system scale	Data-center clusters, hyperscale cloud regions, edge nodes, computing-power networks, and data-center buildings connected to local energy systems.
Included coordination variables	Electricity price, demand-response signal, carbon intensity, grid congestion, renewable availability, cooling state, storage state, and service latency.
Included control actions	Workload deferral, workload migration, server scaling, virtual-machine/container placement, cooling set-point control, battery scheduling, and heat reuse.
Excluded topics	Pure chip architecture, generic cloud load balancing without electricity coupling, cryptocurrency mining economics, and thermal-only studies without workload or grid interaction.
Main survey question	How can electricity and computing power be co-developed so that computation remains reliable while the grid receives measurable flexibility and carbon benefit?

II. Research Scope and Research Challenge

A. Integrated Infrastructure for Electricity and Computing Systems

The classic research on data centers involves lowering facility energy consumption through better cooling, server consolidation, and power-management mechanisms [2], [16], [17]. While these are necessary actions, local efficiency does not imply grid compatibility. Even a very efficient plant, if it sits at an "overloaded node," runs during a period of high demand, and/or draws electricity during carbon-intensive periods, can create strain on the grid. Similarly, a cloud scheduler that tries to minimize computing delay can lead to more emissions if it does not consider the electricity mix and network congestion at the receiving site. The concept of electricity-compute coordination therefore goes beyond keeping electricity costs down at the local level to considering when, where, and how electricity is used to maximise the combined energy, carbon, and service outcome. Electricity-compute coordination introduces a multi-domain problem. From the power-system side, factors to consider are time-varying prices, renewable uncertainty, network constraints, reserve products, carbon accounting, and distribution capacity. On the computing side, these include workloads, queues, service-level agreements (SLAs), data-transfer cost, hardware use, and latency. The added complexities of cooling and storage relate to the fact that a building's power does not move in lockstep with the IT load. This yields an integrated electrical, computational, and thermal system.

B. Literature Review and Categorization Methodology

The literature is organized by primary decision variable rather than publication year. Demand-response studies are grouped under market coupling because the primary action is the response of data centers to electricity-system incentives [4]–[9]. Carbon-aware computing studies are treated separately because their primary objective is emissions reduction rather than price response [3], [10], [11], [22], [23]. Edge-computing and computing-power-network studies are grouped under orchestration because they focus on resource abstraction, routing, and scheduling [13]–[15], [18]–[21], [24]. Cooling, storage, waste heat, and multi-energy operation are grouped under local flexibility because these resources enable a data center to operate as a prosumer or multi-energy hub [16], [17], [25]–[27]. A method is classified as implementation-oriented when it uses measurable signals, satisfies service constraints, can be executed within practical computational time, and does not require unrealistic information sharing among operators. Methods that assume perfectly mobile workloads, complete information across all data centers, deterministic renewable forecasts, or unavailable market mechanisms are classified as simulation-oriented. This distinction is used in the critical analysis because superior simulation results do not necessarily imply greater practical grid value.

C. Low-Latency Carbon-Aware Power and Computing Coordination

Latency-guaranteed, nodal-carbon-aware workload-power co-dispatch is identified as a key research challenge. Operation with latency guarantees requires workload movement to remain within SLAs. For nodal-carbon-aware operation, carbon signals need to be based on local grid conditions and tied to marginal generation rather than only a regional average. Workload placement, server power, cooling power, storage, and grid flexibility should all be co-determined in co-dispatch. The formulation may be restrictive but sufficiently concrete to support a focused research program and bridge the gap between power systems, cloud systems, and edge networks. Current approaches typically address only subsets of this problem. Carbon-aware data-center operation has demonstrated practical value, but carbon signals are often aggregated and decoupled from grid constraints [3]. Demand-response models capture load flexibility but may omit internal computing workflows or latency classes [6]–[9]. Computing-power-network studies describe networked computing-resource orchestration, yet electricity-market clearing and distribution-grid constraints are often absent [13]–[15]. Edge-computing scheduling can reduce latency and energy consumption but generally does not quantify grid-level carbon benefits [18]–[21], [24]. The principal unresolved issue is therefore the absence of a common operational layer for trusted exchange of carbon, grid, and workload information.

III. Classification of Electricity and Computing Coordination Approaches

Table 2. Main taxonomy for electricity-compute coordination literature.

Layer	Main idea	Typical methods	Implementation maturity	Common metrics	Refs.
Electricity-market and demand-response coupling	Treats data centers as controllable loads that respond to prices, incentives, reserves, or grid requests.	Game theory, contract design, optimization, demand-response baselines, load aggregation.	Medium to high: compatible with today's utility programs, but needs accurate baselines and SLA-safe curtailment.	Peak reduction, cost saving, reserve contribution, SLA violation, ramping.	[4]–[9], [12]
Carbon-aware workload and location scheduling	Moves, delays, or throttles computation using carbon intensity, renewable availability, or long-term siting signals.	Carbon-aware dispatch, spatio-temporal workload migration, siting optimization, emissions accounting.	High for batch tasks; medium for latency-sensitive services; low if nodal carbon and network limits are ignored.	Carbon reduction, energy cost, latency, utilization, renewable matching.	[3], [10], [11], [22], [23], [28]–[30]
Computing-power network and edge-cloud orchestration	Advertises and routes computing resources through network-aware scheduling across cloud, edge, and network domains.	Computing resource abstraction, task routing, offloading, service-function placement, edge scheduling.	Medium: architecture is promising, but interoperator standards and electricity-aware routing are still immature.	Task delay, throughput, resource use, communication cost, reliability.	[13]–[15], [18]–[21], [24]

Local energy, cooling, and multi-energy flexibility	Coordinates IT load with cooling systems, batteries, renewable generation, uninterruptible power supply (UPS), heat recovery, or district energy.	Model predictive control (MPC), mixed-integer optimization, thermal modeling, microgrid scheduling, heat-reuse planning.	Medium to high for single sites; lower for distributed portfolios without market interfaces.	Power usage effectiveness (PUE), carbon usage effectiveness (CUE), energy cost, thermal safety, heat revenue, battery cycling.	[16], [17], [25]–[27]
Evaluation, verification, and deployment evidence	Checks whether claimed flexibility is measurable, repeatable, and acceptable to both grid and computing operators.	Field trials, real traces, power-flow simulation, carbon accounting, auditable settlement.	Still weak: many studies use synthetic workloads or ideal forecasts, and few report distribution-grid impacts.	Nodal emissions, congestion relief, baseline error, SLA loss, operator revenue, user delay.	[3], [7], [10]–[12]

A. Electricity-Market and Demand-Response Coupling

Demand-response studies examine whether data centers can modify power consumption in response to electricity-price or grid-reliability signals. Early geo-distributed models assume that services can be delivered from multiple sites, allowing workload to shift from high-price locations to alternative sites without violating quality-of-service requirements. A game-theoretic framework modeled this interaction between geo-distributed data centers and electricity markets [4]. Subsequent studies added contracts and markets to make such flexibility available to utilities [5], [6]. This translates computing flexibility into services for power systems in terms of load reduction, price response, reserve provision, and settlement. One of the biggest issues is creating a credible "baseline" for a demand-response system. A utility requires a credible estimate of the electricity that would have been consumed without a demand-response event. Computing demand is elastic, forecast-dependent, and may be opaque within cloud-provider operations. Baseline errors can therefore lead to over- or under-compensation for flexibility. Another distinction is between demand shifting and actual demand reduction. Deferring a batch job may reduce peak power during an event but create rebound demand later. This can still provide short-term grid relief, but carbon benefits depend on the time and location of the deferred computation. Accordingly, market-coupled models should be evaluated using both electrical and computing metrics rather than cost alone.

A series of recent load models has been developed to enhance deployability by taking into account multiple forms of regulation within a data center, such as server scaling, cooling control, and workload shifting [8]. Strategic behavior by data centers is also a concern in incentive-compatible models. These mechanisms play a crucial role because operators have little incentive to offer flexibility or divulge information about their operations. However, most studies do not consider constraints at the distribution level, feeder capacity to host renewables, or local-transformer limits.

B. Energy-Efficient Workload and Location Management

In carbon-aware computing, the main control signal is emission intensity rather than electricity price. Studies have shown that data-center workloads can be scheduled in low-carbon locations or at low-carbon times without violating data-center operations [3]. This is the principle of making computing work directly related to decarbonization goals. Rather than relying only on renewable certificates or facility-efficiency improvements, carbon-aware scheduling exploits workload flexibility to reduce emissions. Carbon-oriented demand response extends this concept by integrating computing load into broader decarbonization pathways [10], [11]. The effectiveness of these methods depends strongly on the carbon signal used. Average grid carbon intensity is widely available but may not represent the generator responding to incremental demand. Marginal carbon intensity better reflects dispatch effects but is more difficult to measure or predict. In constrained networks, nodal marginal emissions are more informative because they incorporate local power-system conditions, although such data and models are often unavailable to cloud operators. This creates a significant research gap: reported carbon reductions depend strongly on the accounting method. Moving workload to a low-average-carbon region can still increase fossil generation if the receiving grid is congested or renewable generation is already fully utilized. Long-term siting and capacity planning therefore complement real-time dispatch. Carbon-aware location and configuration models identify regions with favorable renewable resources, climate, and grid capacity [23], [28].

Recent studies of electricity-computing synergy also incorporate renewable-resource endowments and geographical constraints [28]. These planning-oriented approaches are valuable for infrastructure siting but do not replace operational dispatch. A well-sited facility can still operate during periods of high local demand or congestion. Effective coordination therefore requires both planning decisions, such as facility location and capacity, and operational decisions, such as hourly or sub-hourly workload migration.

C. Integrated Computing Power and Edge-Cloud Coordination

From a communications perspective, computing-power networks provide the orchestration layer for electricity-compute coordination. Computing resources, such as bandwidth and storage, can be sensed, abstracted, advertised, and routed through a network [13]–[15]. This is relevant to providers seeking to shift computation from demand-constrained regions toward locations with greater renewable availability. It is equally important for edge computing, where tasks must remain sufficiently close to users to satisfy latency requirements. Existing surveys of computation placement primarily evaluate delay, energy consumption, and network-resource utilization [18]–[21], [24], but rarely assess whether the selected computing node also benefits the electrical grid. A nearby node may satisfy latency requirements while using carbon-intensive electricity; a cleaner node may be too distant for latency-sensitive control or video-analytics tasks. Energy-aware edge computing addresses part of this trade-off through server and device energy metrics [18], but carbon intensity, renewable curtailment, and distribution-grid constraints are less frequently modeled. Computing-power-network research can address this gap by incorporating electricity-related attributes into resource advertisements, including available power headroom, carbon intensity, thermal margin, flexibility capacity, and grid-congestion status.

Implementation depends critically on trust and interoperability. There is limited freedom to share internal data about providers' or utilities' operations to support routing or orchestration. Standardized information is thus needed to reveal the necessary coordination data while keeping key business or grid-security information private. For instance, instead of publishing detailed server states, a computing node can publish a flexibility envelope, and instead of publishing detailed congestion and/or carbon signals, a utility could publish the signals in an appropriately aggregated way. In the absence of these interfaces, a computing-power network is not a fully-fledged electricity-compute coordination platform but remains a network-optimization architecture.

D. Energy Storage and Heat Recovery Systems

In addition to servers, a data center is composed of cooling plants, power electronics, UPS systems, batteries, backup power generation, on-site renewable power, and possibly heat-recovery equipment. These assets can provide local flexibility. Cooling set-points are programmable; batteries can store surplus renewable energy and/or reduce peak demand; and waste heat can be provided to district-heating systems if needed at the same time it is produced [25]–[27]. Thermal dynamics, power consumption, and control opportunities differ between air- and liquid-cooling systems [16], [17].

Multi-energy data-center models are particularly suitable for building- or campus-scale implementation because site operators directly control local energy assets. Studies show that integrating renewables, storage, and active demand response can reduce cost and emissions [25]. Data centers can also interact with district-energy systems through waste-heat export [26], [27]. These approaches reduce dependence on remote workload migration because local assets can reshape the facility's grid interaction even when computing tasks cannot move. However, local optimization does not necessarily produce system-wide grid benefits, making scale coupling an important limitation. Coordinated operation across multiple sites can also create rebound effects. If many facilities respond simultaneously to the same low-price period, a new system peak may emerge. Heat-recovery equipment may be underutilized when seasonal heat demand is low, while cooling flexibility must respect hardware temperature and reliability limits. Local flexibility should therefore be coordinated at portfolio and grid levels. The relevant research objective is not to compare cooling and workload shifting in isolation, but to integrate workload, cooling, storage, and heat recovery as a unified flexibility stack.

E. Modeling and Evaluation Metrics

Evaluation is a major source of inconsistency across the literature. Data-center electricity use is commonly represented by IT power, cooling power, and facility-overhead metrics such as PUE. Carbon-aware studies may report emissions reductions using average, marginal, or nodal carbon intensity, while edge-computing studies emphasize latency and throughput, and power-system studies emphasize cost, peak reduction, ramping, and reserve provision. Evaluation should therefore be integrated into the methodology rather than treated as a secondary component. There must be at least four metric groups for a practical assessment. Electrical metrics include peak reduction, shifted energy, ramp smoothing, transformer loading, reserve availability, and power-flow feasibility. Computing measures include latency, queue time, SLA violation rate, dropped requests, throughput, and data-transfer overhead. Examples of environmental measures are energy usage, carbon emissions, renewable-energy matching, water footprint, and heat recovery. Economic measures include electricity cost, market revenues, infrastructure cost, battery degradation, and customer penalties. The removal of any of these groups may result in a partial view of system performance. The most convincing data are actual workload traces, field operations, and/or hardware-in-the-loop testbeds. Combining field evidence with synthetic simulation increases the credibility of carbon-aware operation [3]. Similarly, to better represent the validity of market models, realistic tariffs, demand-response baselines, and operator incentives should be included. Distributions of measured latency and migration overhead [18]–[21], [24] should be part of edge-computing evaluations. Publicly available workload traces with energy prices, carbon intensity, weather, cooling power, facility load, and grid constraints would be useful for future benchmarking. Global datasets of this nature are still rare, however, which means that many modeling assumptions have limited validation. GreenSLA denotes Green Service Level Agreement, and GreenSDA denotes Green Supply Demand Agreement. Table 3 provides an implementation-oriented comparison of the major method families.

Table 3. Implementation-oriented comparison of major method families.

Method family	Main use	Strength	Limitation	Implementation realism	Refs.

Game-theoretic demand response	Coordinates distributed data centers under price or market interaction.	Good for studying incentives and strategic behavior.	Often assumes full rationality and simplified grid constraints.	Medium: useful for market design, weaker for real-time control.	[4], [6], [9]
Contract-based GreenSLA/Green SDA	Defines agreements between cloud services and demand-response programs.	Implementation-friendly because it maps flexibility to contracts.	Needs reliable baselines and customer acceptance of degraded or shifted service.	High for flexible enterprise workloads.	[5]
Carbon-aware workload shifting	Moves or delays work toward cleaner electricity.	Practical for batch and non-urgent AI/cloud workloads.	Carbon signal quality and latency limits determine real benefit.	High for batch jobs; medium for interactive services.	[3], [10], [11]
Bi-level planning/scheduling	Jointly models infrastructure investment and operation.	Captures interaction between computing capacity, storage, and electricity system.	Computationally heavy and dependent on scenario assumptions.	Medium: useful for planning, less for minute-level dispatch.	[28]–[30]
CPN/edge-cloud orchestration	Routes tasks according to computing and network resource status.	Natural architecture for distributed computing services.	Electricity signals and carbon data are not yet standard inputs.	Medium: promising but needs standards.	[13]–[15], [18]–[21], [24]
Cooling/storage/ heat-flexibility control	Uses facility assets to reshape electrical demand.	Deployable because assets are locally controlled.	Thermal safety, battery degradation, and heat demand must be modeled.	High at single-site level; medium at portfolio level.	[16], [17], [25]–[27]
Deep reinforcement learning heuristics	Learns scheduling policies from traces or simulation.	Handles nonlinear systems and many states.	Weak guarantees, data hunger, and poor transfer under rare grid events.	Low to medium unless combined with safety layers.	[18]–[21], [24]

IV. Critical Analysis

The most implementation-ready approaches are those that respect current operational constraints. Carbon-aware workload shifting for batch applications, model training, data analytics, etc., is feasible for secondary tasks in the cloud that have a considerable degree of delay tolerance [3], [10], [11]. Contract-based demand response can also be deployed because flexibility can be codified in a contract or through a utility program [5], [6]. Technically, local control of cooling, storage, and renewable resources is feasible because these resources remain under the control of the data-center operator [16], [17], [25]. These approaches might not provide as much theoretical flexibility as idealized optimization models, but they better reflect existing metering and scheduling practices and business constraints.

Some of these methods rely on simulations, such as large bi-level games, idealized geo-distributed migration models, and edge-cloud scheduling without considering state-transfer costs. Many of these approaches require full workload mobility, perfect forecasts, or complete information. These assumptions can be helpful in establishing a “maximum possible” carbon reduction and grid relief, but achievable benefits may be overstated. Cross-region workload transfer may increase latency, data-transfer energy, and congestion at the receiving node even when average carbon intensity decreases. Likewise, a demand-response simulation could yield peak reductions that do not occur in reality if the baselines are less reliable than assumed or service degradation is not tolerated. Overall, it is evident from Table 3 that none of these method families is sufficient. The market interface is provided by demand response, the environmental goal is defined by carbon-aware scheduling, computing-power networks are used as the routing architecture, and physical flexibility is provided through local multi-energy control. A layered approach is therefore suitable: coordination can be performed day-ahead using market and carbon signals, operational dispatch can use SLA-aware workload classes, and local operation (cooling and storage) can provide rapid adjustments. Approaches providing only synthetic simulations of cost reduction should be avoided in favor of methods that can measure grid impact, SLA performance, and auditable carbon results.

V. Open Challenges and Future Directions

Several challenges remain. First, there is a need for practical coordination based on carbon signals, and these signals must be accurate enough for coordination but sufficiently safe to share at the node level across organisational lines. While regional-average carbon intensity is easy to implement, it might give a misleading picture in heavily loaded networks. Second, workload classes need to be codified so that schedulers can identify delay-tolerant, latency-critical, and location-locked tasks. Third, a reliable demand-response baseline is needed for data centres with elastic, customer-driven compute demand. Fourth, and most importantly, more controlled measurements are needed to define the interactions among server load, temperature, liquid-cooling behavior, storage operation, and battery degradation [16], [17]. Future research should establish testbeds that jointly evaluate real workload traces, electricity prices, weather, carbon signals, and feeder constraints. Settlement is another open issue: when workload migration reduces grid congestion or emissions, the benefit should be measurable and attributable so that participating operators can be compensated. Electricity-aware resource descriptors, such as carbon range, power headroom, and flexibility envelope, should also be integrated into computing-power networks. This would allow resource routing to account for grid conditions in addition to computing performance.

VI. Practical Framework for Integrated Development

Workload classification is the starting point for practical implementation. Data-center operators can partition jobs into four categories: strict-latency, soft-latency, delay-tolerant, and location-locked. Only flexible classes should participate in electricity- and carbon-aware dispatch. The second step is signal integration: electricity prices, grid capacity, renewable forecasts, and carbon signals should be provided at time resolutions appropriate to each workload class. Inference tasks require faster and more conservative decisions, whereas batch tasks can use hourly or day-ahead signals. The third step is local flexibility stacking. Before migrating workload, the facility should evaluate server scaling, cooling set-point adjustment, and battery operation to relieve local constraints. Local actions are typically faster to implement and easier to verify.

The fourth step is portfolio-level coordination across sites. If there are multiple data centers, carbon intensity, electricity price, latency, data-transfer overhead, compute capacity, and grid headroom at the receiving data center(s) should be considered when choosing the workload location. The fifth step is verification. Each flexibility action should be registered in relation to electrical-load impact, computing-service impact, and emissions impact. Verification is required to support grid settlement and credible sustainability reporting. These steps together create a pathway for staged implementation. The first stage is the realization of multi-energy control at a single site. In stage 2, delay-tolerant tasks are introduced with carbon-aware scheduling. In stage 3, markets and carbon pricing are used to coordinate multiple locations. Stage 4 integrates computing-power-network descriptions and introduces electric-power-related attributes into resource routing. Full-system testing at stage 5 includes utilities, cloud operators, and independent carbon accounting. This step-by-step process is more feasible than using a single optimization algorithm that needs unfettered access to confidential operational data owned by each participant.

VII. Evidence and Benchmarking Protocol

Assessment of evidence should be done in stages. Level 1 uses a mathematical model with a well-explained range of assumptions to estimate a theoretical operating range. At Level 2, actual workload traces are used, as well as measured electricity-price, weather, or facility-measurement data. Level 3 connects workload simulation to feeder-capacity, congestion, and marginal-emissions models. Field operation or hardware-in-the-loop testing with auditable metering is used at Level 4. Most current studies focus on Levels 1 and 2; future studies need to increasingly focus on Levels 3 and 4 so that quantifiable grid flexibility can be presented, not only optimized curves. Further benchmarks should also differentiate workload classes because latency requirements vary dramatically depending on the use case. For example, batch training should be distinguished from scientific workflows or industrial-control inference. At a minimum, a benchmark should measure the following: fraction of load that can be shifted, maximum time window allowed for load shifting, amount of data being shifted, capability at the receiving end, and computing penalty. Peak reduction, ramp-rate reduction, carbon reduction, or other relief of grid-node constraints must be explicitly identified as electrical impacts. It should be explicitly stated that average, marginal, and nodal marginal carbon intensities differ. Table 4 outlines the minimum evidence protocol for improved comparison and reduced overclaiming. Lower-evidence studies can still be valuable when explicitly identified as the outcomes of simulation or planning studies. Stronger studies use multiple sources of evidence, such as realistic market and cost parameters, sensitivity analysis under forecast error, real computing workloads, and measured energy/cooling data under real conditions. This assessment can help separate cost-shifting tactics from approaches that deliver verifiable grid services without compromising the quality of computing services.

Table 4. Minimum evidence protocol for future electricity-compute coordination studies.

Evidence level	What it means	Minimum information required	How to interpret the result
Level 1: Model-only study	Optimization or control model with synthetic workload and assumed power system.	Clear mathematical assumptions, constraint list, and sensitivity analysis.	State that the result is a theoretical operating envelope, not deployment proof.
Level 2: Trace-based simulation	Real or public computing traces with measured price, weather, or carbon data.	Workload class, latency tolerance, migration cost, and energy/cooling trace.	Report SLA loss, rebound load, and robustness under forecast error.
Level 3: Grid-coupled validation	Workload simulation linked with power-flow, feeder capacity, nodal carbon, or congestion model.	Electrical node, grid constraint, marginal emission method, and demand-response baseline.	Show grid benefit separately from operator cost saving.
Level 4: Field or hardware-in-the-loop test	Metered facility, cloud cluster, emulator, or utility-facing pilot.	Auditable metering, settlement rule, real SLA impact, and repeatable event records.	Highest evidence level; should become the target for future journal studies.

VIII. Conclusion

This survey examines the co-development of electricity and computing resources at the intersection of power systems and data centers, carbon-aware computing, edge-cloud orchestration, and multi-energy operation. The literature shows that data centres can participate in demand-response schemes and move flexible workloads toward green power, as well as utilise local energy resources and connect to wider computing-power schemes. Fragmentation is the main limitation: while several methods optimize price, carbon, latency, or cooling individually, deployment requires a harmonized and verifiable decision loop. Based on this, the proposed taxonomy classifies the literature into five layers and establishes latency-guaranteed workload co-dispatch with nodal-carbon awareness as a highlighted research challenge. This framework may be used as a basis for future models, testbeds, interoperability standards, and grid- and service-aware digital infrastructures.

References

[1] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, “Recalibrating global data center energy-use estimates,” Science, vol. 367, no. 6481, pp. 984-986, 2020. doi: 10.1126/science.aba3758.

[2] M. Dayarathna, Y. Wen, and R. Fan, “Data center energy consumption modeling: A survey,” IEEE Communications Surveys & Tutorials, vol. 18, no. 1, pp. 732-794, 2016. doi: 10.1109/COMST.2015.2481183.

[3] A. Radovanovic et al., “Carbon-aware computing for datacenters,” IEEE Transactions on Power Systems, vol. 38, no. 2, pp. 1270-1280, 2023. doi: 10.1109/TPWRS.2022.3173250.

[4] N. H. Tran, D. H. Tran, S. Ren, Z. Han, E. N. Huh, and C. S. Hong, “How geo-distributed data centers do demand response: A game-theoretic approach,” IEEE Transactions on Smart Grid, vol. 7, no. 2, pp. 937-947, 2016. doi: 10.1109/TSG.2015.2421286.

[5] R. Basmadjian, J. F. Botero, G. Giuliani, X. Hesselbach, S. Klingert, and H. de Meer, “Making data centers fit for demand response: Introducing GreenSDA and GreenSLA contracts,” IEEE Transactions on Smart Grid, vol. 9, no. 4, pp. 3453-3464, 2018. doi: 10.1109/TSG.2016.2632526.

[6] S. Bahrami, V. W. S. Wong, and J. Huang, “Data center demand response in deregulated electricity markets,” IEEE Transactions on Smart Grid, vol. 10, no. 3, pp. 2820-2832, 2019. doi: 10.1109/TSG.2018.2810830.

[7] M. Chen, C. Gao, M. Song, S. Chen, D. Li, and Q. Liu, “Internet data centers participating in demand response: A comprehensive review,” Renewable and Sustainable Energy Reviews, vol. 117, Art. no. 109466, 2020. doi: 10.1016/j.rser.2019.109466.

[8] M. Chen, C. Gao, M. Shahidehpour, Z. Li, S. Chen, and D. Li, “Internet data center load modeling for demand response considering the coupling of multiple regulation methods,” IEEE Transactions on Smart Grid, vol. 12, no. 3, pp. 2060-2076, 2021. doi: 10.1109/TSG.2020.3048032.

[9] M. Chen, C. Gao, M. Shahidehpour, and Z. Li, “Incentive-compatible demand response for spatially coupled Internet data centers in electricity markets,” IEEE Transactions on Smart Grid, vol. 12, no. 4, pp. 3056-3069, 2021. doi: 10.1109/TSG.2021.3053433.

[10] T. Wan, Y. Tao, J. Qiu, and S. Lai, “Internet data centers participating in electricity network transition considering carbon-oriented demand response,” Applied Energy, vol. 329, Art. no. 120305, 2023. doi: 10.1016/j.apenergy.2022.120305.

[11] B. Du, H. Jia, Y. Li, E. Du, N. Zhang, and D. Liang, “Estimating the carbon emission reduction potential of using carbon-oriented demand response for data centers: A case study in China,” iEnergy, vol. 4, no. 1, pp. 54-64, 2025. doi: 10.23919/IEN.2025.0007.

[12] T. Yang, Y. Hou, S. Cai, J. Yu, and H. Pen, “Multi-data center tie-line power smoothing method based on demand response,” IEEE Transactions on Cloud Computing, vol. 12, no. 4, pp. 983-995, 2024. doi: 10.1109/TCC.2024.3410377.

[13] Y. Sun, B. Lei, J. Liu, H. Huang, X. Zhang, J. Peng, and W. Wang, “Computing power network: A survey,” China Communications, vol. 21, no. 9, pp. 109-145, 2024. doi: 10.23919/JCC.ja.2021-0776.

[14] Q. Zhao, B. Lei, and M. Wei, “Survey of computing power network,” ITU Journal on Future and Evolving Technologies, vol. 3, no. 3, pp. 632-644, 2022. doi: 10.52953/BXBJ6384.

[15] B. Lei, Q. Zhao, and J. Mei, “Computing power network: An interworking architecture of computing and network based on IP extension,” in Proc. 2021 IEEE 22nd International Conference on High Performance Switching and Routing (HPSR), pp. 1-6, 2021. doi: 10.1109/HPSR52026.2021.9481792.

[16] Q. Zhang, Q. Meng, S. Ma, and X. Yang, “A survey on data center cooling systems: Technology, power consumption modeling and control strategy optimization,” Journal of Systems Architecture, vol. 119, Art. no. 102253, 2021. doi: 10.1016/j.sysarc.2021.102253.

[17] A. H. Khalaj, T. Scherer, and S. K. Halgamuge, “A review on efficient thermal management of air- and liquid-cooled data centers: From chip to the cooling system,” Applied Energy, vol. 205, pp. 1165-1188, 2017. doi: 10.1016/j.apenergy.2017.08.037.

[18] C. Jiang, T. Fan, H. Gao, W. Shi, L. Liu, C. Cérin, and J. Wan, “Energy aware edge computing: A survey,” Computer Communications, vol. 151, pp. 556-580, 2020. doi: 10.1016/j.comcom.2020.01.004.

[19] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” IEEE Communications Surveys & Tutorials, vol. 19, no. 4, pp. 2322-2358, 2017. doi: 10.1109/COMST.2017.2745201.

[20] P. Mach and Z. Becvar, “Mobile edge computing: A survey on architecture and computation offloading,” IEEE Communications Surveys & Tutorials, vol. 19, no. 3, pp. 1628-1656, 2017. doi: 10.1109/COMST.2017.2682318.

[21] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, “Convergence of edge computing and deep learning: A comprehensive survey,” IEEE Communications Surveys & Tutorials, vol. 22, no. 2, pp. 869-904, 2020. doi: 10.1109/COMST.2020.2970550.

[22] F. Wang, V. Nian, P. E. Campana, J. Jurasz, H. Li, L. Chen, W.-Q. Tao, and J. Yan, “Do ‘green’ data centres really have zero CO2 emissions?” Sustainable Energy Technologies and Assessments, vol. 53, Art. no. 102769, 2022. doi: 10.1016/j.seta.2022.102769.

[23] F. Wang, C. Lv, and J. Xu, “Carbon awareness oriented data center location and configuration: An integrated optimization method,” Energy, vol. 278, Art. no. 127744, 2023. doi: 10.1016/j.energy.2023.127744.

[24] Q. Luo, S. Hu, C. Li, G. Li, and W. Shi, “Resource scheduling in edge computing: A survey,” IEEE Communications Surveys & Tutorials, vol. 23, no. 4, pp. 2131-2165, 2021. doi: 10.1109/COMST.2021.3106401.

[25] D. Xu, S. Xiang, Z. Bai, J. Wei, and M. Gao, “Optimal multi-energy portfolio towards zero carbon data center buildings in the presence of proactive demand response programs,” Applied Energy, vol. 350, Art. no. 121806, 2023. doi: 10.1016/j.apenergy.2023.121806.

[26] P. Huang, A. Copertaro, X. Zhang, J. Shen, I. Löfgren, M. Rönnelid, J. Fahlen, D. Andersson, and M. Svanfeldt, “A review of data centers as prosumers in district energy systems: Renewable energy integration and waste heat reuse for district heating,” Applied Energy, vol. 258, Art. no. 114109, 2020. doi: 10.1016/j.apenergy.2019.114109.

[27] S. S. Ravi, T. S. Löffler, E. A. Pina, S. Sharma, D. Lepour, C. Terrier, and F. Maréchal, "From servers to services: Modeling data centers as heat-active urban energy prosumers," Applied Energy, vol. 402, Part B, Art. no. 127049, 2026. doi: 10.1016/j.apenergy.2025.127049.
[28] J. Guo, W. Hou, X. Wang, X. Zhang, S. Wu, L. Yang, W. Jiang, and L. Zhang, "Toward a new spatial paradigm of electricity-computing synergy under renewable energy endowments and geographical environmental constraints," Applied Energy, vol. 409, Art. no. 127495, 2026. doi: 10.1016/j.apenergy.2026.127495.
[29] S. Zhang, M. Wei, Y. Li, S. Ding, and Y. Chen, "A bi-level two-stage computing-electric power coupled scheduling model of data center cluster with shared energy storage incorporating non-cooperative and cooperative game," Applied Energy, vol. 415, Art. no. 127930, 2026. doi: 10.1016/j.apenergy.2026.127930.
[30] W. Wen, R. Xie, Q. Tang, Z. Xiong, R. Zhang, G. Xie, S. Sun, and T. Huang, "Demand-driven server deployment for green computing power networks: A multi-objective hierarchical optimization approach," IEEE Transactions on Green Communications and Networking, advance online publication, 2026. doi: 10.1109/TGCN.2026.3653570.