


INNOVATION  
– PRESS –

# Energy Use

Energy Use 3 (2026):100074

Contents lists available at INNOVATION PRESS

## AI Data-Center Flexibility is Conditional, Not Intrinsic

Yuren Yang,<sup>1,2</sup> Yuhao Chen,<sup>3</sup> Mufeng Yuan,<sup>1,2</sup> Hao Zhou,<sup>2,4</sup> Yang Zhao,<sup>1,2</sup> Yang Geng,<sup>1,2</sup> Wei Cheng,<sup>5</sup> Hanzao Feng,<sup>6</sup> Jiaying Wang,<sup>3</sup> and Borong Lin<sup>1,2,\*</sup>
<sup>1</sup>School of Architecture, Tsinghua University, Beijing 100084, China

<sup>2</sup>Key Laboratory of Eco Planning and Green Building, Ministry of Education, Beijing 100084, China

<sup>3</sup>State Grid Zhejiang Marketing Service Center, Hangzhou 310007, China

<sup>4</sup>Institute for Urban Governance and Sustainable Development, Tsinghua University, Beijing 100084, China

<sup>5</sup>China United Network Communications Co., Ltd. Guangdong Branch, Guangzhou 510603, China

<sup>6</sup>China Unicom (Guangdong) Industrial Internet Co., Ltd., Guangzhou 510320, China

\*Correspondence: [linbr@tsinghua.edu.cn](mailto:linbr@tsinghua.edu.cn)

Received: September 20, 2026; Accepted: September 22, 2026; Published Online: September 23, 2026; <https://doi.org/10.59717/ipj.energy-use.2026.100074>

© 2026 The Author(s). This is an open access article under the CC BY-NC-ND license (<https://creativecommons.org/licenses/by/4.0/>).

Citation: Yang Y., Chen Y., Yuan M., et al. (2026). AI Data-Center Flexibility is Conditional, Not Intrinsic. *Energy Use* 2:100074.

### Dear Editor,

Artificial intelligence (AI) data centers (AIDCs) are increasingly expected to become active flexibility resources for power systems. As AI-driven electricity demand grows, workload shifting, power capping, frequency scaling, and geographical migration are being explored to reduce grid stress and better align computing demand with electricity prices, renewable availability, and carbon intensity. Recent field demonstrations have further shown that operational AIDCs can provide measurable demand response while maintaining service requirements. Yet this emerging narrative risks compressing a heterogeneous set of computing behaviors into a single quantity of flexible load. We argue that AIDC flexibility is conditional, not intrinsic. It emerges only when three questions are answered in sequence: Can the workload move? Should it move for the objective at hand? Can many such movements be delivered together without creating new system constraints? These questions correspond to workload-conditioned feasibility, objective-conditioned value, and system-conditioned deliverability.

### INTRODUCTION

The conventional power-system abstraction of a flexible load reduces a complex resource to a small set of grid-facing parameters, such as available power change, response time, duration, and recovery. This abstraction is useful, but artificial intelligence (AI) data centers (AIDCs) add an important layer of internal heterogeneity: their electrical demand is generated by a continuously changing portfolio of computational jobs, services, models, and hardware states. Training, fine-tuning, batch inference, interactive inference, and emerging agentic services may use similar accelerators but operate under fundamentally different deadlines, latency requirements, memory states, and migration costs.

Recent research has established that computing workloads can provide measurable demand-side flexibility.<sup>1</sup> Operational GPU clusters can reduce power for grid support,<sup>2</sup> workload-aware deferral can identify latency-tolerant computing activity for demand response,<sup>3</sup> and large-language-model (LLM) power profiles can be reshaped through coordinated changes in accelerator quantity and frequency.<sup>4</sup> These advances establish an important point: AI loads can be flexible. The next question is more specific and more consequential: under what conditions does that flexibility exist, and when can it become useful to the grid?

Treating flexibility as an intrinsic facility property obscures three distinct filters between computing demand and grid response. A workload must first be technically actionable. The action must then create value for a specific electricity-system objective. Finally, many individually attractive actions must

remain feasible when combined on shared electrical and computing infrastructure. Flexibility should therefore be understood not as a fixed percentage of facility load, but as a capability that progressively narrows from computational possibility to grid deliverability (Figure 1).

### WORKLOAD-CONDITIONED FLEXIBILITY: CAN IT MOVE?

The first condition is technical actionability. AI workloads differ in at least four forms of elasticity: temporal, service, state, and spatial elasticity.

Temporal elasticity describes how far execution can be advanced, delayed, slowed, interrupted, or resumed. Long-deadline training and batch workloads may offer substantial temporal slack, whereas interactive services generally operate within much narrower response windows. Carbon-aware systems have already exploited temporal elasticity in batch workloads by adapting execution to variations in electricity carbon intensity while respecting completion constraints.<sup>5</sup>

Service elasticity describes how much intervention can be tolerated before violating application-level requirements such as latency, throughput, availability, or completion deadlines. Two workloads with similar computational slack can therefore possess very different actionable flexibility if one supports offline processing while the other serves latency-sensitive users.

State elasticity captures whether the computational state associated with a workload can be released, migrated, or reconstructed. This dimension is particularly important for LLM serving. Inference itself is heterogeneous: prefill and decoding exhibit different compute, memory, latency, and power characteristics, motivating phase-specific architectures such as Splitwise and DistServe.<sup>6,7</sup> Meanwhile, memory-management systems such as PageAttention illustrate how model serving is constrained by large and dynamically changing key-value cache states.<sup>8</sup> Low arithmetic utilization alone therefore does not establish that an accelerator can be consolidated, migrated, or powered down.

Spatial elasticity describes whether computation or requests can move across accelerators, clusters, facilities, or regions. Some batch jobs may be highly mobile if data, software environments, and compatible hardware are available elsewhere. Interactive inference may have little temporal elasticity but retain spatial elasticity through request routing, provided latency, bandwidth, model placement, and data-locality requirements permit it.

These four dimensions explain why idle computing and flexible electricity demand are not equivalent. A workload becomes actionable only when a feasible control action changes its electrical demand while keeping computational and service consequences within acceptable bounds. The relevant question is therefore not simply how much accelerator capacity appears underutilized, but which workload states can be altered, for how long, with

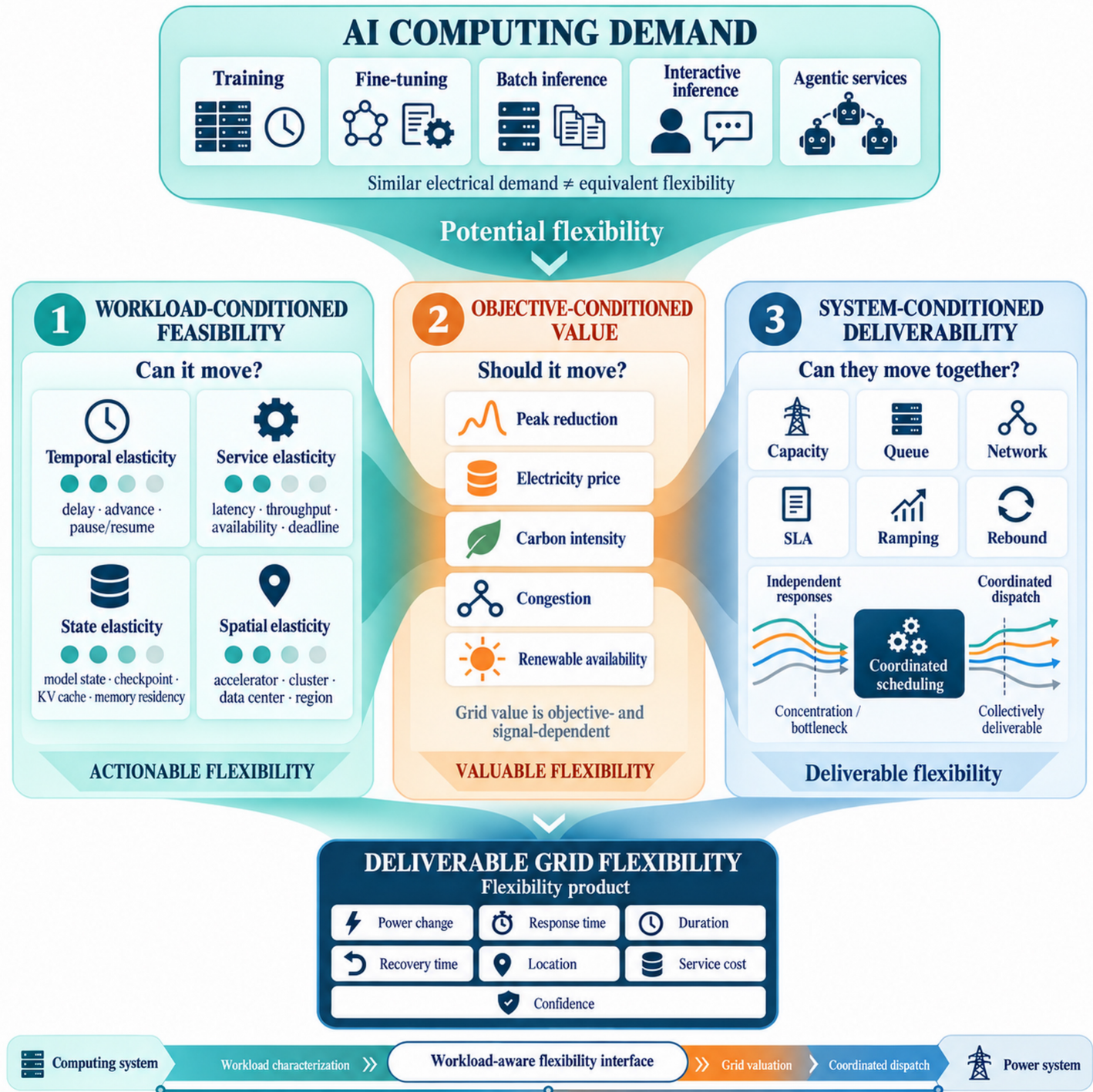


Figure 1. AIDC flexibility emerges through three conditional layers.

what recovery time, and at what service cost.

#### OBJECTIVE-CONDITIONED FLEXIBILITY: SHOULD IT MOVE?

Technical actionability does not imply universal grid value. A workload that can move may be highly valuable for one objective and largely irrelevant for another. Demand response, congestion relief, electricity-cost reduction, renewable-energy absorption, and carbon reduction assign value to different combinations of load magnitude, timing, duration, location, and feasible displacement.

Consider a deferrable batch job. If it overlaps a local system peak, moving it may provide immediate peak-reduction value. If its feasible scheduling window contains little variation in carbon intensity, however, the same job may offer little emissions benefit. Conversely, a long-running workload that does not overlap a demand-response event may still offer substantial carbon

or price value if it can be moved toward a cleaner or cheaper period. The external electricity signal matters as much as the underlying workload.

Existing research already demonstrates this relational character. Carbon-aware scheduling benefits depend on workload elasticity and temporal carbon-intensity variation,<sup>5</sup> while the potential of temporal and geographical workload shifting is constrained by the available scheduling window, regional differences, and application requirements.<sup>9</sup> Demand-response studies similarly show that latency-tolerant workloads become valuable when their feasible delay intersects with a grid event.<sup>3</sup> A workload therefore does not carry a fixed scalar quantity of grid flexibility. Its grid value emerges when a feasible computational action intersects with a particular power-system need.

Grid-facing flexibility can therefore be understood as a relation among three elements: workload attributes and feasible control actions, the grid objective and its associated temporal or spatial signal, and the shared

system context in which those actions must ultimately be coordinated. The first two elements distinguish questions that are often conflated: what can computing do, and what does the grid currently value? Reporting a single facility-level flexibility percentage collapses these questions and can obscure which workloads actually warrant intervention. Computing-electricity coordination should therefore move from identifying generic flexible workloads toward identifying objective-specific high-value workload states.

### SYSTEM-CONDITIONED FLEXIBILITY: CAN THEY MOVE TOGETHER?

Even a technically feasible and individually valuable workload adjustment may fail to become deliverable grid flexibility when many workloads respond simultaneously. Individual flexibility is not simply additive.

AI workloads share electrical feeders, server capacity, queues, networks, cooling infrastructure, and scheduling resources. Price-aware or carbon-aware schedulers may independently direct many jobs toward the same low-price or low-carbon interval. Geographic schedulers may similarly steer requests toward the same favorable region. The result can be a new concentration of electricity demand rather than a reduction in system stress.

This phenomenon has been identified as a stampede effect in carbon-aware cloud computing: independently rational workload decisions can become collectively undesirable when they respond to the same external signal.<sup>10</sup> The underlying implication is broader than carbon-aware scheduling. A set of individually beneficial workload decisions does not automatically constitute a system-optimal schedule.

The distinction is between individual actionability and aggregate deliverability. Once high-value workload actions have been identified, they must be recombined under shared constraints such as facility or feeder capacity, ramping limits, queue dynamics, hardware compatibility, network congestion, service redundancy, and rebound behavior. A low-price or low-carbon interval indicates when electricity consumption may be attractive, but not how much additional flexible demand that interval can safely accommodate.

This third condition changes the role of workload profiling. Profiling should not directly produce a final dispatch schedule. Instead, it should define the feasible action space and prioritize candidate workloads. A subsequent coordination layer must determine which combinations can be executed simultaneously. Computing-electricity coordination therefore requires both fine-grained workload intelligence and aggregate system constraints.

### FROM FLEXIBILITY POTENTIAL TO DELIVERABLE FLEXIBILITY

The three conditional layers imply a progressively narrowing interpretation of AI data-center flexibility. Potential flexibility describes the gross space suggested by nominal computational slack or scheduling freedom. Actionable flexibility retains only workload states that can actually be changed without unacceptable computational or service consequences. Valuable flexibility further retains actions that improve a specified grid objective under the prevailing temporal and spatial signal. Deliverable flexibility is the final subset that can be coordinated at scale while satisfying shared electrical and computing constraints.

A practical computing-electricity coordination architecture should mirror this progression. At the computing layer, workload managers should expose attributes that govern actionability, including deadline slack, interruptibility, latency tolerance, migration capability, state-release requirements, and recovery time. At the grid-valuation layer, feasible actions should be evaluated against explicit objectives such as peak reduction, electricity price, carbon intensity, congestion, or renewable availability. At the coordination layer, selected actions should be jointly scheduled under aggregate capacity and operational constraints.

Importantly, this architecture does not require grid operators to access every internal detail of an AI application. Instead, computing systems could translate workload states into standardized flexibility products characterized by attributes such as available power change, response time, duration, recovery time, location, confidence, and service cost. Different workload classes could then provide fundamentally different products: fast but shallow power modulation, delayed but sustained load shifting, geographically mobile demand, or deeper reductions that require workload consolidation and state release.

The rapid expansion of AI infrastructure makes this distinction increasingly important. Evidence now supports moving beyond the question of whether AI data centers can provide flexibility. The more useful questions are which workloads can respond, to which grid objectives, and under what shared-system constraints. Training, batch inference, interactive inference, and emerging agentic services may all appear as megawatts on the same meter, but they do not represent interchangeable grid resources.

AI data-center flexibility should therefore be treated as a conditional capability rather than an intrinsic facility attribute. Workload properties determine whether a workload can move. Grid objectives determine whether moving it is valuable. System constraints determine whether many such actions can be delivered together. The appropriate unit of computing-electricity coordination is not the aggregate megawatt alone, but the workload state and the conditions under which that state can be translated into a reliable grid service.

### REFERENCES

1. International Energy Agency. (2025). Energy and AI. IEA.
2. Colangelo P., Coskun A. K., Megrue J. et al. (2026). AI data centres as grid-interactive assets. *Nat. Energy* **11**:254–261. DOI:10.1038/s41560-025-01927-1
3. Caprara A., Yu Y., Teng F., et al. (2026). Data center workload flexibility for power system demand response: Evidence from Alibaba traces. *Int. J. Elect. Power Energy Syst.* **178**:111940. DOI:10.1016/j.jepes.2026.111940
4. Wang Y., Guo Q. and Chen M. (2025). Providing load flexibility by reshaping power profiles of large language model workloads. *Adv.Appl. Energy* **19**:100232. DOI:10.1016/j.adapen.2025.100232
5. Hanafy W. A., Liang Q., Bashir N., et al. (2023). CarbonScaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. *Proceed. ACM Measur. Anal. Comput. Syst.* **7**:57. DOI:10.1145/3626788
6. Patel P., Choukse E., Zhang C., et al. (2024). Splitwise: Efficient generative LLM inference using phase splitting. *IEEE ISCA*. DOI:10.1109/ISCA59077.2024.00019.
7. Zhong Y., Liu S., Chen J., et al. (2024). DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. *Proceed. OSDI* :193–210. DOI:10.48550/arXiv.2401.09670.
8. Kwon W., Li Z., Zhuang S., et al. (2023). Efficient memory management for large language model serving with PagedAttention. *Proceed. SOSP*. DOI:10.1145/3600006.3613165.
9. Sukprasert T., Souza A., Bashir N., et al. (2024). On the limitations of carbon-aware temporal and spatial workload shifting in the cloud. *Proceed. EuroSys* :924–941. DOI:10.1145/3627703.3650079.
10. Hanafy W. A., Sukprasert T., Souza A., et al. (2026). Untangling the carbon-cost tradeoffs and stampede effect challenges in cloud computing. *IEEE Trans. Comput.* **75**:2715–2729. DOI:10.1109/TC.2026.3667610

### FUNDING AND ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China (Grant No.52425801, 52394223, 52130803, and 72374121), and the National Key Research and Development Program of China (Grant No. 2024YFE0212900).

### DECLARATION OF INTERESTS

The authors declare no competing interests.