

Computational Systems & Infrastructure

REVIEW | COMMISSIONED SYNTHESIS

Energy-Proportional Scheduling for Cloud and Edge Infrastructure

David Miller - Corresponding author

Abstract

This review examines energy-proportional scheduling across cloud and edge infrastructure. The organizing question is when workloads should move in time or place without degrading reliability and service equity. Ten related scholarly sources are synthesized through a decision-centered framework spanning problem definition, mechanism, measurement, evaluation, implementation, and governance. The review does not invent experiments, pooled estimates, or unreported quantitative results. It instead evaluates the strength and transferability of the available evidence, with particular attention to optimizing reported operational energy while exporting cost or emissions elsewhere. The resulting framework links technical or empirical performance to explicit use conditions and identifies tests that should precede wider adoption in sustainable operation of geographically distributed computing.

Keywords: green computing, cloud scheduling, edge computing, carbon awareness, energy proportionality

Synthesis lens	Principal question
Mechanism	the relation among demand, hardware power states, grid carbon intensity, and network transfer costs
Evaluation	energy, carbon, latency, availability, and rebound effects measured over a complete workload lifecycle
Primary risk	optimizing reported operational energy while exporting cost or emissions elsewhere

1. Introduction

Research on energy-proportional scheduling across cloud and edge infrastructure sits at the boundary between a promising component and a consequential decision. The core question is when workloads should move in time or place without degrading reliability and service equity. Answering it requires more than assembling favorable results. It requires a chain of evidence that makes the problem boundary, mechanism, measurement, comparator, uncertainty, and accountable decision visible to the reader.

This article presents a structured narrative review of ten related publications. Sources were selected for topical relevance, traceable publication metadata, and complementary coverage of technical, empirical, and governance questions. No meta-analytic pooling is attempted because the included studies differ in designs, populations, system boundaries, and outcomes. The purpose is decision-oriented synthesis: to identify what each source can support, where transfer remains uncertain, and which evidence would justify the next stage of use.

The central risk is optimizing reported operational energy while exporting cost or emissions elsewhere. A top-line metric or broad policy label can appear decisive while concealing the conditions that produced it. Accordingly, the synthesis separates observation from interpretation and implementation from validation. This separation is especially important when the same system creates benefits for one user or institution while transferring cost, risk, or administrative burden to another.

2. Evidence Base and Problem Boundary

A defensible review begins by defining the unit of analysis. For the present topic, the relevant boundary includes inputs, institutional or technical context, the mechanism producing an output, the person or system acting on that output, and the downstream outcome. Evidence falls outside the boundary when it measures only an intermediate proxy yet is used to claim an end-to-end benefit.

Katal et al. (2022) addresses "Energy efficiency in cloud computing data centers: a survey on software technologies." In this synthesis, that work is used as a source on problem definition within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Juarez et al. (2016) addresses "Dynamic energy-aware scheduling for parallel task-based application in cloud computing." In this synthesis, that work is used as a source on conceptual boundary within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Baccarelli et al. (2017) addresses "Fog of Everything: Energy-Efficient Networked Computing Architectures, Research Challenges, and a Case Study." In this synthesis, that work is used as a source on measurement within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Together, these works provide complementary entry points, but their conclusions should not be flattened into a single ranking. Differences in data generation, setting, temporal horizon, and outcome definition may be substantively meaningful. A review should therefore preserve those differences and state where analogy, rather than direct replication, connects one domain to another.

3. Mechanism and System Design

The operative mechanism is the relation among demand, hardware power states, grid carbon intensity, and network transfer costs. This formulation discourages component-only reasoning. A model, platform, policy, assay, or infrastructure service acquires practical meaning only when its interfaces and use conditions are specified. The evidence chain should describe what information is available, what transformation occurs, which assumptions make that transformation credible, and who is authorized to act.

Chen et al. (2021) addresses "Adaptive and Efficient Resource Allocation in Cloud Datacenters Using Actor-Critic Deep Reinforcement Learning." In this synthesis, that work is used as a source on system mechanism within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Anderson et al. (2023) addresses "Treehouse: A Case For Carbon-Aware Datacenter Software." In this synthesis, that work is used as a source on representation choice within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Perin et al. (2021) addresses "Towards Sustainable Edge Computing Through Renewable Energy Resources and Online, Distributed and Predictive Scheduling." In this synthesis, that work is used as a source on failure analysis within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

These studies also show why modular claims require interface tests. A component may perform well while the complete pathway fails because inputs drift, users interpret outputs differently, recovery semantics are ambiguous, or institutional incentives change. Design documentation should therefore include not only the intended pathway but also foreseeable deviations, degraded modes, and conditions under which the system should stop or defer.

4. Evaluation and Uncertainty

Evaluation should center on energy, carbon, latency, availability, and rebound effects measured over a complete workload lifecycle. Average performance remains useful, but it should be accompanied by distributions, error categories, relevant subgroups or operating regimes, and uncertainty at the point where a decision changes. Comparators must isolate the value of the proposed addition rather than compare a mature intervention with an intentionally weak baseline.

Alsadie (2024) addresses "Advancements in heuristic task scheduling for IoT applications in fog-cloud computing: challenges and prospects." In this synthesis, that work is used as a source on comparative evaluation within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Khan et al. (2023) addresses "A Review on Task Scheduling Techniques in Cloud and Fog Computing: Taxonomy, Tools, Open Issues, Challenges, and Future Directions." In this synthesis, that work is used as a source on uncertainty within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

External validation should introduce meaningful shifts in setting, time, population, workload, market structure, or environmental conditions. A nominally independent sample can still be too similar to test transfer. Conversely, a failed transfer is scientifically informative when the changed conditions are measured and the mechanism of failure is examined rather than hidden in an aggregate score.

5. Translation and Governance

Translation to sustainable operation of geographically distributed computing depends on more than technical readiness. Decision rights, documentation, maintenance, monitoring, appeal, and recovery are part of the intervention. The governance requirement is carbon-aware objectives, service constraints, transparent accounting boundaries, and operator override. Each control should be connected to a named failure mode and should identify an owner, evidence source, review interval, and escalation path.

Hussain and Beg (2019) addresses "Fog Computing for Internet of Things (IoT)-Aided Smart Grid Architectures." In this synthesis, that work is used as a source on implementation within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

Jiang et al. (2019) addresses "An Edge Computing Platform for Intelligent Operational Monitoring in Internet Data Centers." In this synthesis, that work is used as a source on governance within energy-proportional scheduling across cloud and edge infrastructure. The connection is deliberately bounded: the cited study informs the evidence map, but it does not by itself establish readiness for sustainable operation of geographically distributed computing. That distinction keeps the review close to the published record and prevents an adjacent result from being converted into a broader causal claim.

A credible implementation plan also defines sunset and revalidation conditions. New data, software updates, institutional restructuring, population change, or shifts in incentives can invalidate previously reasonable assumptions. Monitoring should therefore test the validity of the evidence chain, not merely whether a service remains online or a headline metric remains stable.

6. Research Agenda

Future studies should preregister or otherwise predefine the intended decision, principal outcomes, exclusions, and analysis plan when feasible. They should report missingness, sensitivity analyses, negative or null findings, and the cost of errors to differently situated stakeholders. Reproducible artifacts should include sufficient metadata to reconstruct data provenance, model or analytical versions, parameter choices, and the sequence from evidence to conclusion.

Cross-disciplinary work needs a shared object of explanation rather than a collection of adjacent vocabularies. For energy-proportional scheduling across cloud and edge infrastructure, that shared object is the complete decision pathway. Technical, clinical, social, environmental, or economic expertise should each change the design or interpretation of the study. When evidence remains incomplete, the useful output is a bounded hypothesis, a transparent uncertainty statement, or a request for additional measurement.

7. Conclusion

The literature supports a disciplined approach to energy-proportional scheduling across cloud and edge infrastructure: define the decision, preserve system boundaries, test consequential failure modes, connect uncertainty to action, and assign accountable control. The strongest conclusion is not that one method is universally superior. It is that credible use in sustainable operation of geographically distributed computing requires an auditable link from source evidence to design choice, evaluation result, and governed decision.

References

- Alsadie, D. (2024). Advancements in heuristic task scheduling for IoT applications in fog-cloud computing: challenges and prospects. *PeerJ Computer Science*, 10, e2128. <https://doi.org/10.7717/peerj-cs.2128>
- Anderson, T., Belay, A., Chowdhury, M., Cidon, A., & Zhang, I. (2023). Treehouse: A Case For Carbon-Aware Datacenter Software. *ACM SIGEnergy Energy Informatics Review*, 3(3), 64-70. <https://doi.org/10.1145/3630614.3630626>
- Baccarelli, E., Naranjo, P. G. V., Scarpiniti, M., Shojafar, M., & Abawajy, J. H. (2017). Fog of Everything: Energy-Efficient Networked Computing Architectures, Research Challenges, and a Case Study. *IEEE Access*, 5, 9882-9910. <https://doi.org/10.1109/access.2017.2702013>
- Chen, Z., Hu, J., Min, G., Luo, C., & El-Ghazawi, T. (2021). Adaptive and Efficient Resource Allocation in Cloud Datacenters Using Actor-Critic Deep Reinforcement Learning. *IEEE Transactions on Parallel and Distributed Systems*, 33(8), 1911-1923. <https://doi.org/10.1109/tpds.2021.3132422>
- Hussain, M. M., & Beg, M. M. S. (2019). Fog Computing for Internet of Things (IoT)-Aided Smart Grid Architectures. *Big Data and Cognitive Computing*, 3(1), 8. <https://doi.org/10.3390/bdcc3010008>
- Jiang, C., Qiu, Y., Gao, H., Fan, T., Li, K., & Wan, J. (2019). An Edge Computing Platform for Intelligent Operational Monitoring in Internet Data Centers. *IEEE Access*, 7, 133375-133387. <https://doi.org/10.1109/access.2019.2939614>
- Juarez, F., Ejarque, J., & Badia, R. M. (2016). Dynamic energy-aware scheduling for parallel task-based application in cloud computing. *Future Generation Computer Systems*, 78, 257-271. <https://doi.org/10.1016/j.future.2016.06.029>
- Katal, A., Dahiya, S., & Choudhury, T. (2022). Energy efficiency in cloud computing data centers: a survey on software technologies. *Cluster Computing*, 26(3), 1845-1875. <https://doi.org/10.1007/s10586-022-03713-0>
- Khan, Z. A., Aziz, I. A., Osman, N. A. B., & Ullah, I. (2023). A Review on Task Scheduling Techniques in Cloud and Fog Computing: Taxonomy, Tools, Open Issues, Challenges, and Future Directions. *IEEE Access*, 11, 143417-143445. <https://doi.org/10.1109/access.2023.3343877>
- Perin, G., Berno, M., Erseghe, T., & Rossi, M. (2021). Towards Sustainable Edge Computing Through Renewable Energy Resources and Online, Distributed and Predictive Scheduling. *IEEE Transactions on Network and*

Service Management, 19(1), 306-321.
<https://doi.org/10.1109/tism.2021.3112796>