Health-Aware AI Inference Scheduling with Long-Term Fairness Objectives

Pengei Li
pflics@rit.edu
Rochester Institute of Technology
Rochester, US

Yuelin Han
yhan116@ucr.edu
University of California, Riverside
Riverside, US

Adam Wierman
adamw@caltech.edu
California Institute of Technology
Pasadena, US

Shaolei Ren
shaolei@ucr.edu
University of California, Riverside
Riverside, US

Abstract

The rapid growth of AI inference workloads has intensified electricity consumption in data centers, exacerbating air pollution and associated public health harms. Unlike carbon emissions, these health impacts are highly localized and disproportionately borne by frontline communities near fossil-fuel-based power generation. Despite growing interest in carbon-aware computing, public health outcomes remain largely absent from system-level optimization and evaluation. We study health-aware geographical load balancing for AI inference, explicitly treating long-term fairness in public health impacts as a first-class objective. We formulate an online optimization framework that captures operating cost, switching cost, and a horizon-wide fairness cost defined over cumulative, location-dependent health damages. Using real-world inference traces and region-specific health damage data, we empirically quantify the trade-offs between efficiency, stability, and fairness. Our results show that accounting for health-aware fairness can substantially reduce cumulative public health inequities, but also exposes fundamental tensions with traditional efficiency-oriented objectives. These findings highlight the importance of incorporating public health metrics into the evaluation and design of AI infrastructure, and provide evidence-based guidance on the costs and benefits of health-aware AI scheduling.

1 Introduction

The rapid expansion of AI driven services has fueled a global surge in data center deployment. While indispensable to modern computing, data centers are highly energy intensive. For example, a report by the U.S. Department of Energy [13] projects that U.S. data centers could account for up to 12.0% of total national electricity demand in the near future. This escalating energy consumption is expected to generate substantial environmental externalities, including increased carbon emissions and localized air pollution, particularly in regions where electricity generation relies heavily on fossil fuels.

Over the past decade, carbon aware cloud computing has attracted sustained attention from both industry [3, 17] and academia [8, 9]. However, an equally critical dimension, namely the public health impacts of air pollution induced by AI related energy demand, has received comparatively limited consideration. Recent studies indicate that the rapid growth of AI infrastructure is contributing to elevated emissions of criteria air pollutants, such as fine particulate matter ($PM_{2.5}$) and nitrogen oxides (NO_x), primarily through increased utilization of fossil fuel based power plants.

Exposure to these pollutants is strongly associated with adverse health outcomes, including higher rates of asthma, cardiovascular disease, and premature mortality. Importantly, these health burdens are not evenly distributed. While the economic and societal benefits of AI are globally diffuse and largely virtual, the associated health costs are geographically concentrated in so called frontline communities located near power plants and data center hubs. Many of these communities are disproportionately low income or composed of racial and ethnic minorities. The magnitude of this disparity is striking. One recent work[4] estimates that the annual public health costs attributable to U.S. data center operations could exceed $20 billion by 2028, a scale comparable to the total on road vehicle emissions of a state such as California. These externalized costs remain largely absent from prevailing sustainability metrics and corporate reporting practices within the technology sector.

Despite the public health risks posed by AI-driven energy demand, system-level interventions offer meaningful opportunities for mitigation. Health damages from electricity consumption vary significantly across locations and time due to differences in generation mix and renewable availability, creating opportunities to reduce overall pollution exposure by shifting AI inference workloads toward cleaner regions or periods. Such *geographical load balancing* is practically feasible, as major cloud providers already operate geographically distributed data centers and demonstrated how to exploit spatial flexibility to reduce energy costs or carbon emissions[5]. However, extending these practices to explicitly account for public health externalities raises new algorithmic questions.

In particular, mitigating health impacts requires balancing efficiency objectives (e.g., energy cost, and reconfiguration overhead) against fairness considerations that reflect the unequal distribution of pollution-related harms across communities.

Motivated by this gap, we conduct an empirical study of *health-aware geographical load balancing* for AI inference. Our central goal is to understand how inference workloads can be scheduled in a manner that explicitly incorporates public health outcomes. Specifically, we aim to characterize the fundamental trade-off between system efficiency and fairness when health impacts are treated as a first-class objective. While recent work has examined environmental equity with respect to carbon emissions and water consumption [7], the corresponding efficiency-fairness trade-offs for health outcomes remain largely unexplored. This paper takes a first step toward quantifying the efficiency sacrifices required to achieve more equitable public health outcomes in large-scale AI computing systems.

Scope and relationship to prior work. This manuscript builds on empirical observations that were previously included only as supplementary/appendix material in our earlier, primarily theoretical, study [6]. Here we extract that empirical component as a standalone investigation and substantially expand it, with a dedicated and systematic analysis of the *health* dimension, which was not the main focus of the prior paper. Accordingly, the goal of this work is not to introduce new theory, but to provide a careful empirical characterization and discussion that complements the theoretical results.

2 System Model and Cost Structure

We study a discrete-time system indexed by $t = 1, \ldots, T$, where each time slot corresponds to one hour. Since all slots have identical duration, we use power and energy interchangeably. At time t, the system faces an exogenous AI inference workload w_t, which should be instantaneously served without deferral. Let $x_t = (x_{1,t}, \ldots, x_{N,t})$ denote the provisioning decision across N geographically distributed data centers, where $x_{i,t}$ represents the allocated server (or GPU) capacity at location i. Each data center has a capacity limit $x_{i,t} \le M_i$, and the workload conservation constraint requires $\sum_{i=1}^{N} x_{i,t} = w_t$. Each provisioning decision incurs three types of costs: a per-round *operating (hitting) cost*, an inter-temporal *switching cost*, and a long-term *fairness cost* capturing cumulative public health impacts. Below we detail the hitting cost, which accounts for electricity consumption and workload imbalance.

Hitting Cost. We model IT power consumption as a linear function of provisioned capacity, given by $qx_{i,t}$ at data center i. To account for non-IT overhead such as cooling and power delivery, we incorporate the power usage effectiveness (PUE) factor γ_i. Let $p^e_{i,t}$ denote the time-varying electricity price at location i. The resulting electricity expenditure at time t is therefore $\sum_{i=1}^{N} p^e_{i,t}\, \gamma_i\, q\, x_{i,t}$. In addition, we penalize imbalanced workload allocation. Spatial imbalance may overload a subset of data centers while leaving others underutilized. To capture both effects in a unified manner, we introduce a quadratic regularization term and define the imbalance penalty for data center i at time t as $u_1\|x_{i,t}\|^2$. When aggregated over time, this term corresponds to the squared ℓ_2 norm of the capacity trajectory $(x_{i,1}, \ldots, x_{i,T})$, thereby discouraging both spatial skewness and temporal volatility. Specifically, we define hitting cost

$$f_t(x_t) = \sum_{i=1}^{N} p^e_{i,t}\, \gamma_i\, q\, x_{i,t} + u_1 \sum_{i=1}^{N} \|x_{i,t}\|^2, \tag{1}$$

where the first term represents energy cost and the quadratic regularizer penalizes spatially skewed allocations.

Switching Cost. We introduce a switching cost to discourage excessive temporal variation in provisioning decisions:

$$d(x_t, x_{t-1}) = \frac{u_2}{2}\|x_t - x_{t-1}\|^2. \tag{2}$$

This term models the operational overhead associated with hardware reconfiguration and communication when dynamically rerouting AI workloads, with u_2 controlling the severity of such penalties.

Fairness Cost Following the methodology of [16], we characterize the time-varying health damage rate across data center locations by a vector $A_t = (A_{1,t}, \ldots, A_{N,t})$, which reflects pollution-related health risks per unit of electricity consumed. While routing computation toward cleaner locations or time periods can reduce aggregate damage, it is equally important to prevent disproportionate harm to specific regions. To this end, we incorporate a long-term fairness cost that regularizes cumulative public health impacts over the entire horizon:

$$g(x_{1:T}) = u_3 \left\| \frac{q}{T} \sum_{t=1}^{T} A_t\, x_t^\top \right\|_p. \tag{3}$$

The ℓ_p norm provides a flexible mechanism to interpolate between average and worst-case notions of fairness, recovering a min-max objective as $p \to \infty$. By explicitly accounting for horizon-wide health externalities, this formulation promotes equitable risk distribution while preserving flexibility for dynamic workload routing. By summing up the hitting, switching and long-term fairness cost, the objective function is defined as below

Metrics	OPT	FairOPT	HITMIN	ROBD	DMD	FairOBD		
						$\eta = 10^{-2}$	$\eta = 10^{-3}$	$\eta = 10^{-4}$
Hitting Cost	171.30	177.20	**159.75**	163.63	169.46	170.06	167.80	167.47
Switching Cost	23.27	351.48	43.75	**23.16**	93.53	25.65	27.52	28.17
Fairness Cost	36.36	33.33	140.12	111.85	54.16	**41.36**	51.05	52.68
Total Cost	230.93	562.00	343.62	298.64	317.15	**237.07**	246.36	248.31

Table 1. The average costs of different algorithms in the default setting (i.e. $u_1 = 10$, $u_2 = 1000$ and $u_3 = 3.5$). Minimum costs for the online algorithms are highlighted in bold.

$$\begin{aligned} &\arg\min_{x_{1:T}} \frac{1}{T}\sum_{t=1}^{T} f_t(x_t) + \frac{1}{T}\sum_{t=1}^{T} d(x_t, x_{t-1}) + g\Big(x_{1:T}\Big) \\ s.t. \quad &\sum_{i=1}^{N} x_{i,t} = w_t, \quad \forall t \in [1,T] \\ &x_{i,t} \le M_i, \ \forall i \in [1,N], \ \forall t \in [1,T] \end{aligned} \tag{4}$$

Datasets. We use a publicly available inference trace dataset for LLM services on Azure [14], consisting of coding-related inference requests processed by multiple LLM services. The traces are collected between May 10th and May 16th, 2024, and capture user demand patterns across different times of the week. We aggregate requests at an hourly granularity to simulate the total computational workload over the course of a day.

We assume that the total workload can be distributed across seven data centers located in Arizona, Iowa, Illinois, Texas, Virginia, Washington, and Wyoming ($N = 7$), with location information obtained from [10]. Electricity costs are computed using the average state-level industrial electricity price [15], denoted by $p^e_{i,t}$. The Power Usage Effectiveness (PUE) parameter γ_i is taken from [11].

To quantify health impacts, we account for air pollutant emissions induced by electricity consumption, which lead to adverse outcomes such as increased hospitalizations and lost school days. Following standard epidemiological and economic models, these effects are monetized as health costs. WattTime [16] provides region-specific, time-varying health prices ($/MWh), representing the health cost per unit of electricity consumption. Using regional mappings from [10], we estimate the health cost for each data center and set $A_t = [\gamma_i \cdot p^h_{i,t}]_{i=1}^N$, where $p^h_{i,t}$ is the hourly average health price from WattTime.

Baseline Algorithms. We consider the following baseline algorithms for comparison.

- **Optimal Fairness Offline (FairOPT)**: an offline benchmark that minimizes only the long-term fairness cost over the entire horizon, serving as a lower bound for fairness-aware objectives.
- **Hitting Cost Minimizer (HITMIN)**: an online baseline that greedily selects the action minimizing the instantaneous hitting cost at each time step, without considering temporal coupling or long-term fairness.
- **Regularized Online Balanced Descent (ROBD)**: a state-of-the-art online algorithm designed for smoothed online convex optimization with switching costs, which does not incorporate long-term fairness objectives.
- **Dual Mirror Descent (DMD)**: an online algorithm that focuses exclusively on optimizing the long-term fairness objective via dual updates, while ignoring switching costs.
- **Optimal Offline (OPT)**: the strongest offline benchmark with full knowledge of the cost sequence, corresponding to the $(T, 0)$-OPT benchmark in our analysis. No online algorithm can outperform this oracle benchmark.

Among these algorithms, ROBD achieves the lowest competitive ratio for smoothed online convex optimization [2], which only considers hitting and switching costs when scheduling the workloads. Besides, DMD is proposed for online allocation problems focusing on long-term costs [1, 12], including such fairness costs considerations.

3 Empirical Evaluation

We evaluate the proposed algorithm under a realistic load balancing setting motivated by public health–aware data center operations. In this context, the *hitting cost* captures instantaneous operational efficiency, the *switching cost* reflects infrastructural and institutional inertia, and the *fairness cost* quantifies long-term disparities in health impact induced by geographically distributed computing activities.

Overall performance and fairness–efficiency trade-off. Table 1 summarizes the average costs of all algorithms under the default setting. Across a wide range of learning rates η, FairOBD consistently achieves the lowest total cost among all online algorithms, while simultaneously attaining a reasonable hitting cost when compared with OPT. This indicates that explicitly accounting for long-term fairness

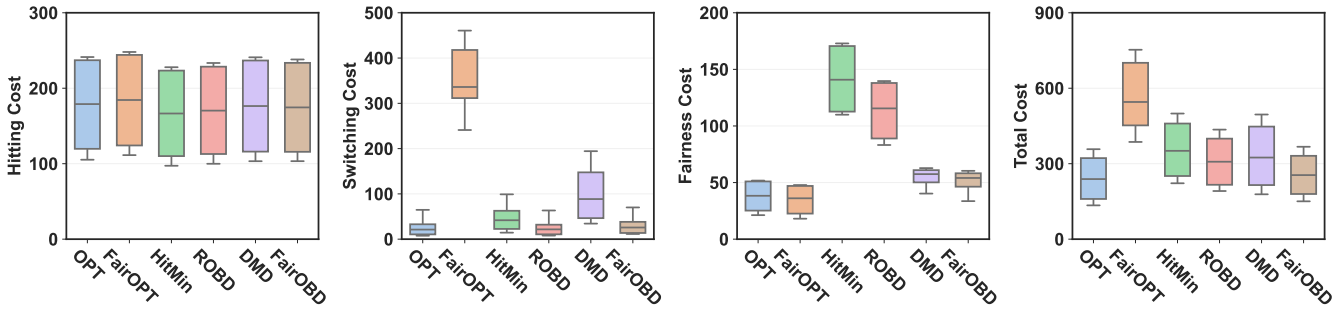

Figure 1. Cost distributions of different algorithms in the default setting (i.e. $u_1 = 10$, $u_2 = 1000$ and $u_3 = 3.5$)

objectives does not come at the expense of operational efficiency, but rather enables a more balanced and socially aligned decision-making process.

Offline baselines and structural trade-offs. To contextualize the achievable fairness performance, we include `FairOPT` as an offline fairness-optimal benchmark with full knowledge of the future. While `FairOPT` attains the lowest fairness cost, it incurs substantially higher hitting and switching costs. This behavior reflects a fundamental tension in public health–aware resource allocation: locations associated with lower health impact do not necessarily coincide with regions offering lower electricity prices, and aggressively shifting workloads toward such locations induces large-scale reconfiguration over time. As a result, `FairOPT` serves as a conceptual lower bound on fairness rather than a deployable policy.

In contrast, `HITMIN` represents the opposite extreme by prioritizing instantaneous operational efficiency. As shown in Figure 1, this strategy leads to significantly increased fairness and switching costs, demonstrating that efficiency-driven policies alone may exacerbate long-term health disparities and induce unstable provisioning behavior.

Limitations of myopic smoothing and the role of long-term structure. The algorithm `DMD` partially mitigates this tension by decomposing long-term objectives and optimizing them via mirror descent, achieving a better balance between hitting and fairness costs. However, without explicitly enforcing action smoothness, `DMD` still incurs relatively high switching costs, which correspond to frequent workload relocation that are undesirable in real-world data center operations.

The `ROBD` algorithm effectively minimizes switching cost by interpolating between the current action and the hitting-cost minimizer. While this yields the lowest switching cost among online baselines, it fails to control the fairness cost. This limitation highlights a key insight: long-term fairness objectives are inherently non-separable and cannot be enforced through local smoothing or per-round balancing alone.

Effectiveness of fairness-aware optimization. By introducing auxiliary variables to represent long-term fairness deviations and dynamically optimizing them via mirror descent, `FairOBD` explicitly captures the temporal structure of fairness costs. This design enables `FairOBD` to achieve the best overall trade-off among hitting, switching, and fairness costs, resulting in the lowest total cost among all online methods.

Distributional stability and worst-case behavior. Beyond average performance, Figure 1 highlights substantial differences in the distributional behavior of competing methods. Algorithms that prioritize either instantaneous efficiency (`HITMIN`) or fairness alone (`FairOPT`) exhibit heavy-tailed cost distributions, particularly in switching and total cost, indicating large variability across time. Such tail behavior corresponds to episodic but severe reconfigurations of server provisioning, which are undesirable from both operational and public health perspectives.

In contrast, `FairOBD` consistently produces tighter cost distributions across all metrics, suggesting improved stability and reduced exposure to extreme outcomes. This property is especially important in health-aware settings, where sporadic but large reallocations may concentrate environmental or health burdens on specific regions. By explicitly regulating long-term fairness while preserving smooth action evolution, `FairOBD` mitigates worst-case behaviors and promotes more predictable and equitable system dynamics.

4 Conclusion

This work provides the first systematic empirical characterization of health-aware AI inference scheduling with long-term fairness objectives. Our results reveal a fundamental and previously underexplored insight: public health fairness is inherently a horizon-wide, non-separable objective that cannot be enforced through myopic efficiency optimization or local smoothing alone. Approaches that focus solely on instantaneous cost minimization or short-term stability consistently lead to concentrated and inequitable health burdens, even when overall system efficiency appears acceptable.

References

[1] Santiago Balseiro, Haihao Lu, and Vahab Mirrokni. 2021. Regularized Online Allocation Problems: Fairness and Beyond. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 630–639. https://proceedings.mlr.press/v139/balseiro21a.html

[2] Gautam Goel, Yiheng Lin, Haoyuan Sun, and Adam Wierman. 2019. Beyond online balanced descent: An optimal algorithm for smoothed online optimization. *Advances in Neural Information Processing Systems* 32 (2019), 1875–1885.

[3] Udit Gupta, Young Geun Kim, Sylvia Lee, Jordan Tse, Hsien-Hsin S. Lee, Gu-Yeon Wei, David Brooks, and Carole-Jean Wu. 2022. Chasing Carbon: The Elusive Environmental Footprint of Computing. *IEEE Micro* 42, 4 (jul 2022), 37–47. doi:10.1109/MM.2022.3163226

[4] Yuelin Han, Zhifeng Wu, Pengfei Li, Adam Wierman, and Shaolei Ren. 2024. The unpaid toll: Quantifying the public health impact of ai. *arXiv preprint arXiv:2412.06288* (2024).

[5] Ross Koningstein. 2021. We now do more computing where there's cleaner energy. Google Blog: The Keyword. https://blog.google/company-news/outreach-and-initiatives/sustainability/carbon-aware-computing-location/ Accessed: January 31, 2026.

[6] Pengfei Li, Yuelin Han, Adam Wierman, and Shaolei Ren. 2025. Fairness-Regularized Online Optimization with Switching Costs. *Advances in Neural Information Processing Systems* (2025).

[7] Pengfei Li, Jianyi Yang, Adam Wierman, and Shaolei Ren. 2024. Towards Environmentally Equitable AI via Geographical Load Balancing. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems* (Singapore, Singapore) *(e-Energy '24)*. Association for Computing Machinery, New York, NY, USA, 291–307. doi:10.1145/3632775.3661938

[8] Zhenhua Liu, Yuan Chen, Cullen Bash, Adam Wierman, Daniel Gmach, Zhikui Wang, Manish Marwah, and Chris Hyser. 2012. Renewable and cooling aware workload management for sustainable data centers. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems* (London, England, UK) *(SIGMETRICS '12)*. Association for Computing Machinery, New York, NY, USA, 175–186. doi:10.1145/2254756.2254779

[9] Zhenhua Liu, Minghong Lin, Adam Wierman, Steven Low, and Lachlan L. H. Andrew. 2015. Greening Geographical Load Balancing. *IEEE/ACM Transactions on Networking* 23, 2 (2015), 657–671. doi:10.1109/TNET.2014.2308295

[10] Microsoft. [n. d.]. Microsoft Local Communities. https://local.microsoft.com/communities/.

[11] Microsoft. 2023. Microsoft Data Centers Sustainability - Efficiency. https://datacenters.microsoft.com/sustainability/efficiency/.

[12] Michael J Neely. 2010. Universal scheduling for networks with arbitrary traffic, channels, and mobility. In *49th IEEE Conference on Decision and Control (CDC)*. IEEE, 1822–1829.

[13] A. Shehabi, S. J. Smith, N. Horner, I. Azevedo, R. Brown, J. Koomey, E. Masanet, D. Sartor, M. Herrlin, and W. Lintner. 2016. United States Data Center Energy Usage Report. *Lawrence Berkeley National Laboratory, Berkeley, California. LBNL-1005775* (2016).

[14] Jovan Stojkovic, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. 2025. Dynamollm: Designing llm inference clusters for performance and energy efficiency. In *2025 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 1348–1362.

[15] U.S. Energy Information Administration (EIA). 2024. EIA OPEN DATA. https://www.eia.gov/opendata/.

[16] WattTime. 2024. Data signals for WattTime dataset. https://watttime.org/data-science/data-signals/

[17] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable AI: Environmental Implications, Challenges and Opportunities. In *Proceedings of Machine Learning and Systems*, Vol. 4. 795–813.