European International Journal of Multidisciplinary Research and Management Studies (ISSN: 2750-8587)

Received: 01 January 2026 | Accepted: 15 January 2026 | Published: 31 January 2026

RESEARCH ARTICLE

Proxy-Oriented Thermal and Acoustic Intelligence for Cloud GPU Orchestration in AI-Guided Scientific Workflows

Marek Zielinski

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

VOLUME: Vol.06 Issue 01 2026

PAGE: 114-123

Abstract

The accelerating convergence of artificial intelligence, cloud computing, and large-scale scientific simulation has fundamentally altered the operational, architectural, and epistemological foundations of high-performance computing. Contemporary workloads ranging from genome-scale language modeling and molecular dynamics to physics-informed reservoir simulation and data-driven optimization increasingly rely on cloud-hosted graphical processing units whose performance, reliability, and sustainability are mediated not only by digital metrics such as throughput and latency but also by physical phenomena such as thermal dissipation, acoustic vibration, and hardware-induced stochasticity. This article advances a comprehensive theoretical and methodological framework for understanding how proxy-based thermal and acoustic evaluation of cloud GPUs can be systematically integrated into AI-driven scientific workflows to improve computational efficiency, reliability, and scientific validity. Building on the seminal contribution of Lulla, Chandra, and Sirigiri, who demonstrated that thermal and acoustic proxies offer predictive insight into the performance and degradation patterns of cloud GPUs under AI training loads (Lulla et al., 2025), this study situates hardware-aware proxies within a broader ecosystem of task-based execution frameworks, data movement systems, surrogate modeling, and AI-guided simulation pipelines.

The paper synthesizes literature from high-performance computing, cloud services, molecular simulation, generative AI for materials and biology, and proxy-based optimization in engineering domains to argue that physical proxies represent an underexplored but theoretically powerful layer of observability for distributed AI workloads. Through an extended conceptual methodology grounded in heterogeneous computing theory, streaming AI–HPC coupling, and task-based performance modeling, the article proposes a multi-layer proxy architecture in which thermal and acoustic signals are interpreted as latent variables reflecting computational stress, scheduling inefficiency, and data movement bottlenecks. The Results section offers a detailed interpretive analysis of how such proxies can be aligned with existing performance suites, workflow engines, and AI-driven adaptive simulations to enable anticipatory scheduling, energy-aware orchestration, and fault-tolerant execution, drawing on studies of ProxyStore, TaPS, Globus Compute, and AI-guided biomolecular and materials design (Pauloski et al., 2024; Ward et al., 2023; Zvyagin et al., 2023).

KEYWORDS

Cloud GPU performance, thermal and acoustic proxies, AI-guided simulation, proxy-based modeling, heterogeneous computing, scientific workflows

INTRODUCTION

The contemporary landscape of scientific computing is increasingly defined by the convergence of artificial

intelligence, cloud-native infrastructure, and high-performance simulation, a convergence that has transformed both the scale and the epistemic character of computational science. In domains as diverse as protein folding, genome-scale language modeling, molecular dynamics, and subsurface reservoir optimization, the computational substrate is no longer a monolithic supercomputer but a heterogeneous assemblage of cloud GPUs, data services, serverless functions, and AI-driven orchestration layers that dynamically adapt to workload demands (Ward et al., 2023). This transformation has generated unprecedented opportunities for accelerating discovery, but it has also introduced new layers of complexity, opacity, and physical vulnerability that challenge traditional models of performance evaluation and system reliability. Within this context, the notion that thermal and acoustic signals emitted by cloud GPUs can serve as meaningful proxies for computational state, efficiency, and degradation represents a profound shift in how we conceptualize observability in distributed AI systems, a shift powerfully articulated by Lulla et al. (2025).

Historically, performance analysis in high-performance computing has relied on digital metrics such as floating-point operations per second, memory bandwidth, network latency, and I/O throughput. These metrics, while indispensable, are abstractions that deliberately ignore the physical realities of computation, including heat generation, mechanical vibration, and power fluctuation. In classical supercomputing environments, where hardware is tightly controlled and workloads are relatively homogeneous, this abstraction has been justified by the stability and predictability of the physical substrate. In contrast, cloud GPUs supporting AI training and simulation workloads operate in highly variable environments, where multi-tenancy, dynamic provisioning, and aggressive power management create complex thermal and acoustic patterns that directly affect performance and reliability (Lulla et al., 2025). The recognition that these physical patterns can be captured and interpreted as proxies for computational behavior challenges the longstanding separation between physical and logical layers of computing.

This reconceptualization resonates strongly with broader trends in scientific modeling, where proxy-based and surrogate approaches have emerged as powerful tools for managing complexity. In petroleum engineering, for example, neural network proxies and physics-informed surrogates have been widely adopted to approximate reservoir behavior, enabling rapid well placement optimization and production forecasting without the prohibitive cost of full-physics simulation (Mohaghegh, 2022; Kolajoobi et al., 2023). Similarly, in molecular science, AI-driven surrogate models have been used to guide adaptive simulations, dramatically accelerating protein folding and materials discovery (Brace et al., 2022; Park et al., 2024). These developments suggest a deeper epistemological shift toward indirect, proxy-mediated representations of complex systems, a shift that now extends to the computational infrastructure itself.

The work of Lulla et al. (2025) represents a pivotal moment in this trajectory by demonstrating that thermal and acoustic emissions from cloud GPUs can be systematically correlated with AI training workloads, revealing patterns of utilization, stress, and inefficiency that are invisible to conventional performance counters. By framing these physical signals as proxies, their study implicitly aligns hardware monitoring with the broader tradition of proxy modeling in science and engineering, suggesting that cloud infrastructure can be treated as a complex system whose internal states must be inferred rather than directly observed. This perspective is particularly salient in the era of AI-guided workflows, where adaptive scheduling, dynamic data movement, and real-time model steering depend on accurate and timely knowledge of system state (Pauloski et al., 2024).

At the same time, the rise of task-based execution frameworks and cloud-native workflow systems has created an unprecedented need for fine-grained, cross-layer observability. Systems such as TaPS, Dask, ProxyStore, and Globus Compute enable the decomposition of large scientific problems into thousands or millions of tasks distributed across heterogeneous resources (Pauloski et al., 2024; Bauer et al., 2024). While these frameworks offer remarkable scalability and flexibility, they also introduce new sources of performance variability and failure that cannot be fully captured by traditional metrics. Thermal throttling, power capping, and hardware aging can silently degrade GPU performance, leading to unpredictable slowdowns, numerical instability, and even silent data corruption, phenomena that are increasingly relevant for long-running AI training and simulation workflows (Lulla et al., 2025).

The introduction of thermal and acoustic proxies into this ecosystem raises profound theoretical and practical questions.

From a theoretical standpoint, it challenges the prevailing view that performance is a purely digital phenomenon, suggesting instead that computation must be understood as a cyber-physical process in which information processing and energy dissipation are inseparably intertwined. From a practical standpoint, it invites the development of new orchestration strategies that treat physical signals as inputs to scheduling, load balancing, and fault tolerance algorithms. These questions are particularly urgent in the context of AI-guided scientific discovery, where the cost of computational error or inefficiency can propagate into scientific conclusions, as in the case of mispredicted protein structures or suboptimal reservoir development strategies (Zvyagin et al., 2023; Qi et al., 2023).

Despite the promise of proxy-based hardware observability, the literature remains fragmented. Studies of cloud services for AI-guided simulation emphasize data movement, task scheduling, and workflow composition but rarely consider physical-layer metrics (Ward et al., 2023). Conversely, work on thermal and acoustic monitoring has traditionally focused on reliability engineering and data center management rather than scientific workflow optimization (Lulla et al., 2025). Meanwhile, the extensive body of research on proxy and surrogate modeling in engineering and the natural sciences has not yet been systematically connected to the problem of computational infrastructure itself (Mohaghegh, 2022; Tang and Durlofsky, 2024). This conceptual gap motivates the present study.

The central argument of this article is that proxy-based thermal and acoustic evaluation of cloud GPUs should be integrated into the design and operation of AI-driven scientific workflows as a foundational layer of observability and control. By synthesizing insights from cloud computing, AI-guided simulation, and proxy modeling across disciplines, the paper seeks to establish a unified theoretical framework in which physical hardware signals are treated as latent variables that mediate the relationship between computational demand and scientific output. This framework is not merely descriptive but normative, implying that future AI–HPC systems should be architected to sense, interpret, and respond to their own physical state in real time.

To develop this argument, the article draws on a wide range of literature. The emergence of genome-scale language models and generative AI for molecular and materials design demonstrates the growing dependence of scientific discovery on GPU-intensive AI workloads (Dharuman et al., 2023; Zvyagin et al., 2023; Park et al., 2024). Studies of adaptive simulation frameworks such as DeepDriveMD and streaming AI–HPC coupling illustrate how AI models can be integrated into simulation loops to steer computation based on intermediate results (Lee et al., 2019; Brace et al., 2022). Cloud services and data movement platforms such as Globus and ProxyStore reveal the infrastructural complexity underlying these workflows (Foster, 2011; Chard et al., 2014; Pauloski et al., 2024). Finally, the rich tradition of proxy modeling in petroleum engineering and reservoir simulation provides a conceptual template for understanding how indirect observables can guide optimization in complex systems (Mohaghegh, 2022; Kolajoobi et al., 2023; Zhuang et al., 2024).

By weaving these strands together, the article aims to articulate a comprehensive vision of physically aware, proxy-driven cloud computing for AI-guided science. The Methodology section elaborates a conceptual and analytical framework for integrating thermal and acoustic proxies into task-based execution and workflow orchestration. The Results section interprets how such integration would manifest in performance, reliability, and scientific throughput, grounded in the existing literature. The Discussion critically evaluates the theoretical implications, limitations, and future directions of this approach, situating it within broader debates about AI infrastructure, surrogate modeling, and the sustainability of large-scale computation.

Throughout, the work of Lulla et al. (2025) serves as a foundational reference point, not only because it provides empirical evidence for the validity of thermal and acoustic proxies but also because it exemplifies a new way of thinking about cloud GPUs as observable, physically grounded entities rather than abstract compute units. By extending and generalizing this perspective, the present study seeks to contribute to a more holistic and scientifically grounded understanding of AI-enabled cloud computing.

METHODOLOGY

The methodological foundation of this study is necessarily theoretical and integrative, reflecting the fact that the problem of proxy-based thermal and acoustic evaluation of cloud GPUs sits at the intersection of multiple disciplinary traditions that are rarely combined in a single empirical protocol. Rather than proposing a narrowly defined experimental design, the

methodology adopted here is a structured analytical synthesis that draws on established practices in high-performance computing performance evaluation, AI-guided simulation, and proxy modeling in engineering and the natural sciences. This approach is justified by the complexity of the phenomena under investigation and by the need to articulate a coherent conceptual framework before large-scale empirical validation can be meaningfully pursued, a position that is consistent with the literature on complex system modeling and surrogate-based optimization (Mohaghegh, 2022; Kolajoobi et al., 2023).

At the core of the methodology lies the concept of a proxy as an indirect observable that captures salient aspects of a system's internal state without requiring full access to its underlying dynamics. In reservoir engineering, proxies such as neural network surrogates or physics-informed models are trained to map input parameters to production outcomes, enabling rapid evaluation of alternative well placements and development strategies (Alpak and Jain, 2021; Qi et al., 2023). In molecular science, AI surrogates approximate free energy landscapes and folding pathways, allowing adaptive simulation frameworks to allocate computational effort where it is most needed (Lee et al., 2019; Brace et al., 2022). In the context of cloud GPUs, Lulla et al. (2025) demonstrate that thermal and acoustic signals emitted during AI training workloads can be treated as proxies for utilization, stress, and potential degradation, suggesting that similar surrogate logic can be applied to computational infrastructure.

The first methodological step, therefore, is to define the proxy space for cloud GPU operation. Drawing on Lulla et al. (2025), thermal proxies are conceptualized as time-varying temperature distributions across GPU components, while acoustic proxies are defined as frequency and amplitude patterns in the audible and ultrasonic emissions generated by fans, power regulators, and mechanical structures. These proxies are not assumed to be perfectly correlated with performance, but they are hypothesized to encode latent variables such as power draw, clock throttling, and mechanical wear that directly affect computational throughput and reliability. This hypothesis is consistent with the broader literature on hardware monitoring and with the empirical correlations reported by Lulla et al. (2025).

The second methodological step is to embed these proxies within a task-based execution and workflow orchestration framework. Modern scientific workflows are increasingly decomposed into fine-grained tasks managed by systems such as Dask, Globus Compute, and custom HPC–AI coupling frameworks (Pauloski et al., 2024; Bauer et al., 2024; Brace et al., 2022). These systems already collect extensive metadata about task execution, including start times, completion times, data transfer volumes, and resource assignments. By extending this metadata model to include thermal and acoustic proxy streams, one can conceptually construct a multi-dimensional observability layer in which each task is associated not only with digital performance metrics but also with physical state indicators. This integration is analogous to the way in which proxy models in reservoir engineering incorporate both static geological features and dynamic production data to inform optimization (Kolajoobi et al., 2023).

The third methodological step involves defining interpretive mappings between proxy signals and actionable system states. In AI-guided simulation, adaptive steering relies on surrogate models that map simulation outputs to decisions about where to allocate further computation (Lee et al., 2019; Park et al., 2024). By analogy, a proxy-based GPU monitoring system would require models that map thermal and acoustic patterns to predictions about performance degradation, imminent throttling, or failure risk. These models could be conceptualized as machine learning surrogates trained on historical correlations between proxy signals and observed performance outcomes, an approach that parallels the use of denoising autoencoders and graph neural networks in subsurface flow optimization and reservoir modeling (Qi et al., 2023; Tang and Durlofsky, 2024).

The fourth methodological step is to articulate how these proxy-informed predictions would feed back into workflow orchestration. In task-based systems, scheduling and load balancing decisions determine which GPU executes which task at what time, and data movement systems such as ProxyStore, Redis, and KeyDB mediate the flow of inputs and outputs (Pauloski et al., 2024; Redis, 2023; Snap Inc., 2023). A proxy-aware scheduler would treat thermal and acoustic indicators as constraints or cost functions, preferentially assigning compute-intensive tasks to GPUs that exhibit lower thermal stress or more stable acoustic signatures. This logic mirrors the way in which reservoir development strategies are optimized by balancing production potential against operational risk using surrogate models (Zhuang et al., 2024;

Musayev et al., 2023).

The fifth methodological step is to situate this proxy-based orchestration within the broader context of AI-guided scientific workflows. In biomolecular simulation, for example, adaptive frameworks such as DeepDriveMD and streaming AI–HPC ensembles dynamically launch new simulations based on intermediate AI predictions of folding pathways (Lee et al., 2019; Brace et al., 2022). In such settings, GPU availability and performance directly influence the fidelity and timeliness of scientific inference. By incorporating thermal and acoustic proxies, the workflow could avoid assigning critical simulations to GPUs at risk of throttling or failure, thereby improving both computational efficiency and scientific reliability, a possibility that aligns with the hardware-aware perspective advanced by Lulla et al. (2025).

The methodological rationale for this integrative framework is grounded in the theory of complex adaptive systems. Cloud-based AI–HPC infrastructures exhibit nonlinear interactions between workloads, hardware states, and environmental conditions, making them difficult to model with purely analytic approaches. Proxy-based models, by contrast, offer a pragmatic way to capture emergent behavior through indirect observation, a strategy that has proven effective in fields ranging from fluid mechanics to reservoir engineering (Mendez, 2023; Mohaghegh, 2022). By extending this strategy to the physical layer of cloud GPUs, the methodology acknowledges that perfect observability is unattainable but that useful control can still be achieved through well-chosen surrogates.

At the same time, the methodology explicitly recognizes its limitations. Thermal and acoustic proxies are influenced by external factors such as ambient temperature, data center cooling policies, and mechanical noise, which can confound their interpretation (Lulla et al., 2025). Moreover, the integration of physical-layer data into cloud orchestration raises practical challenges related to data collection, privacy, and standardization that are not yet fully addressed by existing platforms (Ward et al., 2023). These limitations underscore the need for cautious, theory-driven development rather than naive deployment.

In summary, the methodology of this study consists of a structured conceptual synthesis that maps the logic of proxy modeling from scientific domains such as reservoir engineering and molecular simulation onto the problem of cloud GPU observability. By defining thermal and acoustic signals as proxies, embedding them in task-based execution frameworks, and articulating their role in adaptive orchestration, the methodology provides a coherent basis for interpreting and extending the empirical findings of Lulla et al. (2025) within the broader ecosystem of AI-guided scientific computing.

RESULTS

The results of applying the proxy-based framework outlined above are necessarily interpretive and synthetic, as they draw on a wide body of existing empirical and theoretical work rather than on a single experimental dataset. Nevertheless, when the insights of Lulla et al. (2025) are combined with the literature on AI-guided workflows, task-based execution, and surrogate modeling, a coherent set of outcomes emerges that illuminates the potential impact of thermal and acoustic proxies on cloud GPU orchestration and scientific productivity.

One of the most salient results is the recognition that thermal and acoustic proxies provide a layer of observability that complements and, in some cases, surpasses traditional digital performance metrics. Lulla et al. (2025) demonstrate that fluctuations in GPU temperature and acoustic emissions correlate with changes in workload intensity, power draw, and potential throttling during AI training. When interpreted within a task-based execution framework such as TaPS or Dask, these correlations imply that proxy signals can reveal micro-level performance variations that are not captured by aggregate metrics such as average utilization or task completion time (Pauloski et al., 2024). This finding aligns with the broader observation in performance evaluation that fine-grained, real-time metrics are essential for optimizing heterogeneous systems (Pauloski et al., 2024; Wang et al., 2021).

A second result concerns the potential of proxy-based monitoring to enhance fault tolerance and reliability in AI-guided scientific workflows. In adaptive molecular simulation frameworks such as DeepDriveMD, the failure or slowdown of a single GPU can disrupt the feedback loop between AI models and simulation ensembles, leading to wasted computation or biased sampling (Lee et al., 2019; Brace et al., 2022). By contrast, a proxy-aware orchestration system could detect abnormal thermal or acoustic patterns indicative of impending failure or throttling and proactively reassign tasks, thereby preserving the continuity of the adaptive workflow. This

interpretive result is consistent with the reliability engineering perspective implicit in Lulla et al. (2025), who argue that physical proxies can serve as early warning signals of hardware stress.

A third result emerges when proxy-based GPU monitoring is considered in the context of data-intensive workflows enabled by platforms such as Globus, ProxyStore, and ADIOS 2. These systems are designed to move large volumes of scientific data efficiently across heterogeneous resources (Foster, 2011; Chard et al., 2014; Godoy et al., 2020; Pauloski et al., 2024). However, data movement itself contributes to GPU load and thermal stress, particularly when GPUs are used for preprocessing or in-situ analysis. Thermal and acoustic proxies, therefore, can be interpreted as indirect measures of data movement intensity, allowing orchestration systems to balance computation and I/O more effectively. This interpretation resonates with the literature on integrated reservoir and surface optimization, where proxy models are used to balance subsurface production and surface facility constraints (Kohler et al., 2020; Zhuang et al., 2024).

A fourth result relates to the scalability of AI-guided generative models for science. Genome-scale language models for protein and viral evolution, as well as diffusion models for materials design, require enormous amounts of GPU time (Dharuman et al., 2023; Zvyagin et al., 2023; Park et al., 2024). The cost and energy footprint of these computations are major concerns, and Lulla et al. (2025) suggest that thermal proxies can be used to infer energy efficiency and hardware utilization. When integrated into cloud services that span heterogeneous resources, such proxies could guide the placement of energy-intensive tasks on GPUs that operate within optimal thermal regimes, thereby improving overall sustainability, a goal that is increasingly emphasized in the cloud services literature (Ward et al., 2023).

A fifth result is the conceptual alignment between proxy-based GPU monitoring and the broader trend toward surrogate-driven optimization in engineering and the natural sciences. In well placement and reservoir development, surrogate models enable rapid exploration of design spaces that would be intractable with full-physics simulation (Alpak and Jain, 2021; Kolajoobi et al., 2023; Zhuang et al., 2024). By treating GPU thermal and acoustic signals as proxies for performance and reliability, cloud orchestration can similarly navigate a high-dimensional resource allocation space without exhaustive measurement of every underlying variable. This parallel suggests that the epistemic logic of proxy modeling is transferable across domains, reinforcing the generality of the approach advocated by Lulla et al. (2025).

Finally, a sixth result concerns the potential for new forms of performance evaluation and benchmarking. Traditional suites such as TaPS focus on digital task execution metrics (Pauloski et al., 2024), but the inclusion of thermal and acoustic proxies would enable a more holistic assessment of system behavior that includes physical stress and environmental impact. This expanded notion of performance is particularly relevant for long-running AI training and simulation workflows, where cumulative thermal stress can affect both hardware lifespan and scientific reproducibility, an issue implicitly raised by Lulla et al. (2025).

Taken together, these results suggest that proxy-based thermal and acoustic evaluation is not a marginal diagnostic tool but a transformative lens through which cloud GPU performance, reliability, and sustainability can be understood. By embedding physical proxies into AI-guided workflows, cloud computing can move toward a more adaptive, resilient, and scientifically grounded mode of operation.

DISCUSSION

The integration of proxy-based thermal and acoustic evaluation into cloud GPU orchestration raises a series of deep theoretical, methodological, and practical questions that extend far beyond the immediate context of performance monitoring. At stake is nothing less than the conceptual boundary between computation and its physical substrate, a boundary that has traditionally been treated as sharp but is increasingly revealed to be porous in the era of large-scale AI and cloud-native science. The work of Lulla et al. (2025) provides a crucial empirical anchor for this debate by demonstrating that physical signals emitted by GPUs are not merely byproducts of computation but meaningful indicators of computational state, and the broader literature on proxy modeling and AI-guided workflows suggests that this insight has far-reaching implications.

From a theoretical perspective, proxy-based hardware observability challenges the abstractionist paradigm that has dominated computer science and high-performance computing for decades. In this paradigm, hardware is treated as a black box that executes instructions according to well-

defined digital rules, and performance is measured in terms of instruction counts, memory accesses, and network messages. Thermal and acoustic phenomena are relegated to the domain of data center engineering, separate from the concerns of algorithm designers and workflow architects. Lulla et al. (2025) disrupt this separation by showing that thermal and acoustic proxies carry information about the very workloads that algorithms and workflows generate, implying that physical-layer observability is intrinsically linked to computational semantics.

This reconceptualization resonates with developments in AI-guided simulation, where the boundary between model and computation has already become blurred. In frameworks such as DeepDriveMD and streaming AI–HPC ensembles, AI models do not merely consume simulation outputs; they actively shape the computational trajectory by deciding which simulations to run next (Lee et al., 2019; Brace et al., 2022). In such a setting, the performance and reliability of the underlying GPUs become part of the scientific process, as they influence which parts of the state space are explored and with what fidelity. Thermal throttling or hardware-induced noise can therefore propagate into scientific inferences, a possibility that underscores the importance of physically aware orchestration as advocated by Lulla et al. (2025).

The analogy with proxy modeling in reservoir engineering and subsurface optimization further illuminates this point. In that domain, it is widely accepted that one cannot directly observe all the relevant properties of a reservoir, so one relies on proxies and surrogates to infer behavior from limited data (Mohaghegh, 2022; Kolajoobi et al., 2023). These proxies are not perfect, but they are indispensable for decision-making under uncertainty. By treating cloud GPUs as complex systems whose internal states must be inferred from indirect signals, the proxy-based approach extends this epistemic logic to computational infrastructure itself. This extension suggests that the design of AI–HPC systems should explicitly account for uncertainty and indirect observability at the hardware level, rather than assuming idealized, noise-free execution.

However, this perspective also invites critical scrutiny. One potential counter-argument is that thermal and acoustic proxies are too noisy and context-dependent to support reliable orchestration decisions. Data center environments vary widely in cooling efficiency, ambient temperature, and mechanical design, which could confound proxy signals and lead to false positives or negatives in failure prediction (Lulla et al., 2025). Moreover, cloud providers may not expose fine-grained physical sensor data to users, limiting the practical applicability of proxy-based monitoring. These concerns echo debates in the reservoir engineering literature about the robustness and generalizability of surrogate models, which can fail when applied outside the domain of their training data (Qi et al., 2023; Tang and Durlofsky, 2024).

In response to these critiques, proponents of proxy-based GPU monitoring can point to the success of adaptive and transfer learning techniques in other domains. In well placement optimization, for example, transfer learning and denoising autoencoders have been used to adapt surrogate models across different reservoirs and data regimes (Qi et al., 2023). Similarly, in molecular science, AI models trained on one class of proteins can be fine-tuned to others (Zvyagin et al., 2023). These techniques suggest that proxy models for thermal and acoustic signals could be calibrated and adapted across different hardware and data center contexts, mitigating some of the variability concerns raised by critics.

Another important theoretical issue concerns the relationship between proxy-based hardware observability and the design of cloud services and middleware. Platforms such as Globus, ProxyStore, Kafka, and ZeroMQ are built around the abstraction of data and task streams, not physical hardware states (Foster, 2011; Chard et al., 2014; Hintjens, 2013; Apache Kafka, 2024). Integrating thermal and acoustic proxies into these platforms would require new interfaces, data models, and scheduling algorithms that treat physical signals as first-class inputs. This raises questions about standardization, interoperability, and governance that parallel those faced by the developers of data movement and function-as-a-service platforms (Bauer et al., 2024; Copik et al., 2022).

At a deeper level, the proxy-based approach invites reflection on the sustainability and ethics of large-scale AI computation. The energy consumption and environmental impact of training genome-scale language models and diffusion models for materials design are increasingly recognized as significant concerns (Dharuman et al., 2023; Park et al., 2024). Thermal proxies, as emphasized by Lulla et al. (2025), provide a window into energy efficiency and waste heat generation, which could be used to optimize workloads not only for speed but also for carbon footprint. This possibility aligns with broader efforts to develop green AI and sustainable

computing, suggesting that proxy-based monitoring could play a role in aligning scientific ambition with environmental responsibility.

The discussion would be incomplete without considering the implications for scientific reproducibility and epistemic trust. In computational science, reproducibility depends not only on code and data but also on the stability of the computational environment. If GPU performance varies due to thermal stress or hardware degradation, then identical workflows may produce subtly different results over time, particularly in stochastic AI training and simulation processes (Lulla et al., 2025). By providing a record of physical proxy signals alongside digital execution logs, proxy-based monitoring could enhance the transparency and auditability of scientific workflows, a goal that resonates with the data management and provenance literature (Godoy et al., 2020; Bauer et al., 2024).

Finally, the proxy-based paradigm invites a rethinking of the co-design of hardware, software, and scientific applications. Traditionally, co-design efforts have focused on aligning algorithmic structures with hardware capabilities, such as memory hierarchies and network topologies. The inclusion of thermal and acoustic proxies suggests that co-design should also consider the physical dynamics of hardware operation, such as heat dissipation and mechanical stability. This expanded notion of co-design could lead to new classes of AI accelerators and data center architectures optimized not only for peak performance but for stable, observable, and sustainable operation under AI-guided workloads, a vision that extends the insights of Lulla et al. (2025) into the future of computing.

In sum, the theoretical and practical implications of proxy-based thermal and acoustic evaluation are profound. By bridging the gap between physical hardware states and AI-driven scientific workflows, this approach challenges entrenched abstractions, offers new avenues for optimization and sustainability, and raises important questions about observability, reliability, and trust in computational science. While significant challenges remain, the convergence of evidence from cloud GPU monitoring, AI-guided simulation, and proxy modeling across disciplines suggests that the integration of physical proxies into cloud computing is not only feasible but conceptually compelling.

Conclusion

This article has argued that proxy-based thermal and acoustic evaluation of cloud GPUs represents a transformative approach to understanding and orchestrating AI-driven scientific workloads in heterogeneous cloud environments. Building on the foundational insights of Lulla et al. (2025), who demonstrated the predictive power of physical proxies for GPU performance and stress under AI training loads, the study has situated hardware observability within a broader intellectual tradition of proxy and surrogate modeling that spans molecular science, reservoir engineering, and high-performance computing. By synthesizing literature on AI-guided simulation, task-based execution frameworks, cloud services, and data-driven optimization, the paper has articulated a unified conceptual framework in which physical-layer signals are treated as latent variables that mediate the relationship between computational demand and scientific output.

The analysis suggests that thermal and acoustic proxies can enhance performance evaluation, fault tolerance, sustainability, and reproducibility in AI-guided workflows, particularly in data- and compute-intensive domains such as genome-scale modeling, molecular dynamics, and materials design. At the same time, it has acknowledged the methodological and practical challenges associated with proxy-based monitoring, including noise, variability, and the need for new middleware abstractions. By framing these challenges within the established theory and practice of surrogate modeling, the article has shown that they are not insurmountable but rather part of a familiar landscape of uncertainty and approximation.

Ultimately, the integration of physical proxies into cloud GPU orchestration invites a more holistic and scientifically grounded view of computation, one that recognizes the inseparability of information processing and energy dissipation. As AI continues to reshape the practice of science, such a view will be essential for ensuring that the infrastructure of discovery is as intelligent, adaptive, and trustworthy as the models it supports.

REFERENCES

1. Apache Kafka. 2024. https://kafka.apache.org/. Accessed Feb 2024.
2. Pauloski, J. G., Rydzy, K., Hayot-Sasson, V., Foster, I., and Chard, K. 2024. Accelerating Python Applications with

Dask and ProxyStore. https://arxiv.org/abs/2410.12092.

3. Alpak, F. O., and Jain, V. 2021. Support-Vector Regression Accelerated Well Location Optimization: Algorithm, Validation, and Field Testing. Computational Geosciences, 25, 2033–2054.
4. Lulla, K. L., Chandra, R., and Sirigiri, K. S. 2025. Proxy-based thermal and acoustic evaluation of cloud GPUs for AI training workloads. The American Journal of Applied Sciences, 7(7), 111–127. https://doi.org/10.37547/tajas/Volume07Issue07-12
5. Godoy, W. F., Podhorszki, N., Wang, R., Atkins, C., Eisenhauer, G., Gu, J., Davis, P., Choi, J., Germaschewski, K., Huck, K., Huebl, A., Kim, M., Kress, J., Kurc, T., Liu, Q., Logan, J., Mehta, K., Ostrouchov, G., Parashar, M., Poeschel, F., Pugmire, D., Suchyta, E., Takahashi, K., Thompson, N., Tsutsumi, S., Wan, L., Wolf, M., Wu, K., and Klasky, S. 2020. ADIOS 2: The Adaptable Input Output System. A framework for high-performance data management. SoftwareX, 12, 100561.
6. Dharuman, G., Ward, L., Ma, H., Setty, P. V., Gokdemir, O., Foreman, S., Emani, M., Hippe, K., Brace, A., Keipert, K., Gibbs, T., Foster, I., Anandkumar, A., Vishwanath, V., and Ramanathan, A. 2023. Protein generation via genome-scale language models with bio-physical scoring. Proceedings of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis.
7. Kolajoobi, R. A., Niri, M. E., Amini, S., and Haghshenas, Y. 2023. A Data-Driven Proxy Modeling Approach Adapted to Well Placement Optimization Problem. Journal of Energy Resources Technology, 145, 013401.
8. Lee, H., Turilli, M., Jha, S., Bhowmik, D., Ma, H., and Ramanathan, A. 2019. DeepDriveMD: Deep-learning driven adaptive molecular simulations for protein folding. IEEE/ACM Third Workshop on Deep Learning on Supercomputers.
9. Brace, A., Yakushin, I., Ma, H., Trifan, A., Munson, T., Foster, I., Ramanathan, A., Lee, H., Turilli, M., and Jha, S. 2022. Coupling streaming AI and HPC ensembles to achieve 100–1000x faster biomolecular simulations. IEEE International Parallel and Distributed Processing Symposium.
10. Mohaghegh, S. D. 2022. Smart Proxy Modeling. CRC Press, Boca Raton, FL.
11. Zvyagin, M., Brace, A., Hippe, K., Deng, Y., Zhang, B., Bohorquez, C. O., Clyde, A., Kale, B., Perez-Rivera, D., Ma, H., et al. 2023. GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. The International Journal of High Performance Computing Applications, 37(6), 683–705.
12. Ward, L., Pauloski, J. G., Hayot-Sasson, V., Chard, R., Babuji, Y., Sivaraman, G., Choudhury, S., Chard, K., Thakur, R., and Foster, I. 2023. Cloud services enable efficient AI-guided simulation workflows across heterogeneous resources. Heterogeneity in Computing Workshop.
13. Pauloski, J. G., Hayot-Sasson, V., Gonthier, M., Hudson, N., Pan, H., Zhou, S., Foster, I., and Chard, K. 2024. TaPS: A Performance Evaluation Suite for Task-based Execution Frameworks. IEEE 20th International Conference on e-Science.
14. Park, H., Yan, X., Zhu, R., Huerta, E. A., Chaudhuri, S., Cooper, D., Foster, I., and Tajkhorshid, E. 2024. A generative artificial intelligence framework based on a molecular diffusion model for the design of metal-organic frameworks for carbon capture. Communications Chemistry, 7(1).
15. Foster, I. 2011. Globus Online: Accelerating and democratizing science through cloud-based services. IEEE Internet Computing, 15(3), 70–73.
16. Chard, K., Tuecke, S., and Foster, I. 2014. Efficient and secure transfer, synchronization, and sharing of big data. IEEE Cloud Computing, 1(3), 46–55.
17. Bauer, A., Pan, H., Chard, R., Babuji, Y., Bryan, J., Tiwari, D., Foster, I., and Chard, K. 2024. The Globus Compute Dataset: An open function-as-a-service dataset from the edge to the cloud. Future Generation Computer Systems, 153, 558–574.
18. Tang, H., and Durlofsky, L. J. 2024. Graph Network Surrogate Model for Subsurface Flow Optimization. Journal of Computational Physics, 512, 113132.
19. Qi, J., Liu, Y., Ju, Y., Zhang, K., Liu, L., Liu, Y., Xue, X., Zhang, L., Zhang, H., Wang, H., et al. 2023. A Transfer Learning Framework for Well Placement Optimization

Based on Denoising Autoencoder. Geoenergy Science and Engineering, 222, 211446.

20. Zhuang, X., Wang, W., Su, Y., Yan, B., Li, Y., Li, L., and Hao, Y. 2024. Multi-Objective Optimization of Reservoir Development Strategy with Hybrid Artificial Intelligence Method. Expert Systems with Applications, 241, 122707.

21. Mendez, M. A. 2023. Linear and Nonlinear Dimensionality Reduction from Fluid Mechanics to Machine Learning. Measurement Science and Technology, 34, 042001.

22. Hintjens, P. 2013. ZeroMQ: Messaging for Many Applications. O'Reilly Media.

23. Redis. 2023. https://redis.io/. Accessed Mar 2023.

24. Snap Inc. 2023. KeyDB: A database built for scale. https://github.com/Snapchat/KeyDB. Accessed Mar 2023.

25. Copik, M., Bohringer, R., Calotoiu, A., and Hoefler, T. 2022. FMI: Fast and Cheap Message Passing for Serverless Functions. Scalable Parallel Computing Laboratory, ETH Zurich.