RESEARCH ARTICLE

Optimizing Cloud-Native Lakehouse Architectures for Real-Time Semiconductor Analytics: Balancing Performance, Cost, and Energy Efficiency

Min Yin[1,*] **and Ledee-FI Frank**[1]

[1] University of California, Berkeley, CA 94720, United States

Abstract

This paper presents a cloud-native Lakehouse architecture designed for real-time semiconductor analytics, with a focus on optimizing storage tiering, data lineage, and cost-energy co-optimization. As semiconductor data analytics require processing massive amounts of real-time data, traditional data warehouses are often insufficient in addressing the need for low-latency, high-concurrency queries. The proposed framework leverages cloud-native technologies, such as AWS, Azure, and distributed databases like Apache Doris, to design a dynamic multi-tier storage system that segregates data based on access frequency and volatility, incorporating columnar compression techniques for efficient storage and faster query performance. Moreover, the paper introduces an innovative query routing mechanism and data lineage tracking framework to ensure data transparency and verifiability, critical for industrial applications. A novel cost-energy optimization model is proposed, which balances query performance with resource efficiency, aiming to minimize both operational costs and energy consumption while meeting business KPIs and SLAs. Experimental results demonstrate significant improvements in storage efficiency, query performance, and cost-energy trade-offs, particularly in semiconductor industry use cases. This research offers practical solutions to overcome the limitations of existing architectures and provides valuable insights for the future development of cloud-native platforms for real-time industrial analytics. Further investigation is needed to explore the scalability and adaptability of the proposed model in different data-intensive domains.

Keywords: cloud-native Lakehouse, semiconductor analytics, storage tiering, columnar compression, query routing, cost-energy optimization.

1 Introduction

Real-time analytics in semiconductor manufacturing has shifted from an auxiliary capability to an operational backbone, where equipment telemetry, process control signals, and wafer-lot provenance form a continuously evolving corpus that must be

Submitted: 10 November 2025
Revised: 26 December 2025
Accepted: 12 March 2026
Published: 05 August 2026

Vol. 2, **No.** 3, 2026.
10.62762/TACS.2025.879079

***Corresponding author:**
✉ Min Yin
gmiayinc@gmail.com

Citation
Yin, M., & Frank, L. F. (2026). Optimizing Cloud-Native Lakehouse Architectures for Real-Time Semiconductor Analytics: Balancing Performance, Cost, and Energy Efficiency. *ICCK Transactions on Advanced Computing and Systems*, 2(3), 255–271.

queried under explicit service targets for freshness, tail latency, and auditability [1]. This paper investigates a cloud-native Lakehouse design that treats storage hierarchy, physical layout, cross-engine query routing, and verifiable lineage as a single optimization surface, while aligning the design with tangible business constraints such as key performance indicators, service-level agreements—where tail latency rather than mean latency governs user-perceived quality [2]—and budgetary ceilings on both cost and energy [3]. Considering the above factors, we adopt an evidence-driven stance: the architecture is articulated with implementable mechanisms, the evaluation reconstructs realistic load profiles, and the discussion acknowledges uncertainty where measurement conventions and carbon-intensity estimates may vary across regions and time.

1.1 Motivation and Context

Semiconductor data streams exhibit pronounced temperature skew, heterogeneous schemas that span time-series signals and high-cardinality identifiers, and governance requirements that demand end-to-end provenance and reproducible recomputation. Cloud-native Lakehouse systems, which decouple compute from elastic object storage and exploit columnar formats with vectorized execution, provide a plausible route to lower cost-to-serve and operational agility [4–8]. The open question, to some extent unresolved in prior reports, is whether a production-grade design can simultaneously stabilize the latency tail, preserve freshness under bursty ingest, and offer verifiable lineage while keeping dollars per query and kilowatt-hours per query within acceptable ranges. Learned query optimizers, which replace static cost heuristics with models refined from execution feedback, have demonstrated the value of such data-driven adaptation, further motivating online optimization under changing workload and resource conditions [9].

1.2 Problem Statement and Scope

We study a practical configuration where object storage is the persistence backbone and commodity cluster compute is orchestrated by Kubernetes with SQL engines such as Apache Doris, ClickHouse, and SparkSQL. The system must satisfy explicit targets on freshness windows and P95 and P99 latency for interactive diagnostics, while supporting batch yield analysis and near-real-time alerting. The architectural question is framed as an integrated control-plane problem: a temperature-aware tiering policy, a columnar layout strategy with compression and materialization, a tail-aware cross-engine router, and a lineage model that is verifiable and audit-ready. All optimization decisions are observable through telemetry and cloud billing interfaces, enabling reconciliation of monetary cost with node-level power readings and supporting online adaptation under nonstationary workload conditions [10]. Some elements remain approximate, for example carbon-intensity attribution, which suggests that further research is needed on region-specific factors and temporal variability.

1.3 Research Questions and Design Principles

Our investigation is guided by four questions that connect mechanisms to measurable outcomes. First, whether dynamic tiering driven by access temperature, freshness thresholds, and recomputation cost can reduce total cost while holding P99 latency stable under mixed workloads. Second, how much additional benefit arises when columnar compression and materialized layout are co-designed with routing that explicitly models tail risk and inter-tier fetch penalties. Third, in which ways verifiable lineage, combined with data-quality signals, shortens audit and recomputation paths and informs both placement and routing. Fourth, whether a performance–cost–energy Pareto frontier emerges under explicit KPI and SLA constraints, and how an online optimizer can track it as prices, loads, and carbon intensity shift. These questions translate into design principles: workload awareness rather than static heuristics, feedback loops that learn from observed latency and quality violations, and governance artifacts that are first-class inputs to the optimizer, not post-hoc annotations.

1.4 Contributions

This paper contributes four artifacts implemented on a widely available cloud stack and evaluated with de-identified semiconductor traces. TempAware-Tiering defines migration and remigration rules from a temperature model that integrates access frequency, freshness requirements, and recomputation cost, with safeguards that mitigate thrashing. A tail-aware router, RouteX, extends the cost model beyond average latency to scanned bytes, I/O paths, CPU cycles, and concurrency, and it updates its estimates from telemetry [11, 12]. VeriLineage provides event-level provenance with versioned schemas and hash-based verification, enabling accelerated audit queries and principled replay. CoOpt-C&E formulates a multi-objective optimizer

that minimizes dollars per thousand queries and kilowatt-hours per query under KPI and SLA constraints, and it reports sensitivity to price and carbon-intensity scenarios. The evaluation protocol, which distinguishes cold from warm starts and reconciles billing records with power readings, aims to produce results that are reproducible and, to some extent, generalizable beyond a single deployment.

2 Related Work

Cloud-native Lakehouse systems emerged to reconcile the elasticity and low unit cost of object storage with the transactional guarantees and optimizer sophistication expected of analytical databases. Prior studies have clarified that a unified catalog with snapshot-isolation style transactions mitigates metadata drift and enables schema evolution with reduced downtime, while partitioning and sorting strategies interact strongly with vectorized execution to control scan amplification under selective predicates. These results, although persuasive, often rely on microbenchmarks or synthetic TPC-like mixes that underrepresent high-cardinality identifiers and bursty ingest typical of semiconductor manufacturing. Considering these factors, the evidence base for strict tail-latency objectives under mixed interactive and batch loads remains partial and, to some extent, context dependent.

Work on storage tiering has progressed from static hot–cold separation to policies that react to recency and frequency signals [13]. Such heuristics improve average read cost on object storage backends and reduce cache churn on SSD tiers. In semiconductor manufacturing, moreover, predictive-maintenance and excursion-detection practices operate predominantly on recent equipment telemetry and lot histories [14], producing precisely the skewed, time-decaying access patterns that tiering policies must exploit. Yet most formulations optimize a single proxy for temperature and omit two variables that are critical in industrial settings: the freshness window that bounds acceptable staleness for diagnostics, and the recomputation cost that quantifies the price of materialized view refresh or replay. Studies on columnar compression and physical layout have shown that dictionary and run-length encodings, combined with carefully chosen partition and sort keys, can drastically reduce scanned bytes and CPU cycles. The operational trade-offs, however, are often underarticulated: write amplification induced by aggressive materialization, refresh cadence that drifts from business rhythms, and compaction backlogs that surface precisely when load is most volatile.

Cross-engine query optimization and routing constitute a complementary thread. Vectorized MPP engines exhibit strengths on selective aggregations and wide scans, whereas general-purpose SQL frameworks provide flexibility for heterogeneous operators and fallback execution. Prior art typically adopts static cost models that approximate data-access penalties by average throughput or single-tier latencies. Under high concurrency and tiered storage, cost surfaces become non-convex and tail risk dominates user experience. Cross-engine mapping frameworks such as Musketeer [15] demonstrate the feasibility of decoupling workflow specification from execution engines, yet they do not model inter-tier fetch penalties, cache residency, or concurrency interference, and few systems combine these signals into a tail-aware router that updates its beliefs from live telemetry. The consequence is a gap between planned and realized performance, particularly for the P95 and P99 percentiles that drive service-level adherence.

Governance and lineage have been addressed through metadata graphs that capture source–transform–product relationships and support audit queries and backward tracing. The literature emphasizes model expressiveness and integration with workflow orchestrators, yet practical deployments frequently fragment lineage across ingestion tools, ETL frameworks, and visualization layers. Surveys of provenance systems document a broad design space in which integrity verification is typically best-effort, cryptographic attestability is rarely a first-class goal, and recomputation paths are not always aligned with storage placement or routing decisions; provenance surveys further emphasize that the appropriate form and granularity of provenance depend on its intended use, implying that static, one-size-fits-all governance rules rarely suffice and richer, purpose-driven signals are needed [16]. As a result, lineage exists as a retrospective record rather than an active signal that can steer optimization toward trustworthy and reproducible outputs [17].

Cost- and energy-aware analytics adds another dimension. Recent reports propose reconciling cloud billing data with cluster telemetry to estimate dollars per query and, in some cases, kilowatt-hours per workload. Carbon-intensity modeling is often scenario based, using regional averages that may not reflect temporal variation at sub-hour granularity. Measurement conventions are heterogeneous, and

online adaptation of policies to price spikes or carbon signals is rarely demonstrated beyond offline what-if analyses. Further research is needed to establish robust methodologies that withstand cross-region comparisons and operational noise while remaining lightweight enough for production use.

Synthesizing these strands, the dominant pattern is specialization along single axes: tiering tuned for frequency, layout tuned for scan reduction, routing tuned for average cost, lineage tuned for post hoc audit, and cost–energy tuned for offline planning. An integrated control plane that treats tiering, layout, routing, and verifiable lineage as mutually reinforcing levers, while optimizing for a performance–cost–energy Pareto frontier under explicit KPI and SLA constraints, is less explored. The present work positions itself at this intersection. It adopts temperature-aware tiering that embeds freshness and recomputation cost, co-designs columnar layout with a tail-aware router that learns from telemetry, operationalizes verifiable lineage as an input to placement and recomputation, and introduces a multi-objective optimizer that reports sensitivity to price and carbon-intensity scenarios. By grounding the evaluation in de-identified semiconductor traces and by distinguishing cold from warm starts, the study aims to complement prior laboratory-style evidence with results that are reproducible and, to some extent, transferrable to adjacent industrial analytics domains.

3 Methodology and Framework

This section details the comprehensive methodology that underpins the Lakehouse control plane, which integrates various components for data storage, query execution, and governance. In this framework, we focus on how temperature-driven tiering, layout optimizations, query routing, lineage verification, and energy/cost attributions contribute to a system designed for both performance and sustainability. The design choices are informed by the characteristics of semiconductor electrical measurement data—high-dimensional, device-level records of the kind curated in publicly available open datasets [21]—which stress storage layout and query paths in ways generic benchmarks do not. These elements work together to create a robust, adaptive, and reproducible pipeline, ensuring that service-level objectives (SLOs) are met with an emphasis on both performance and resource consumption.

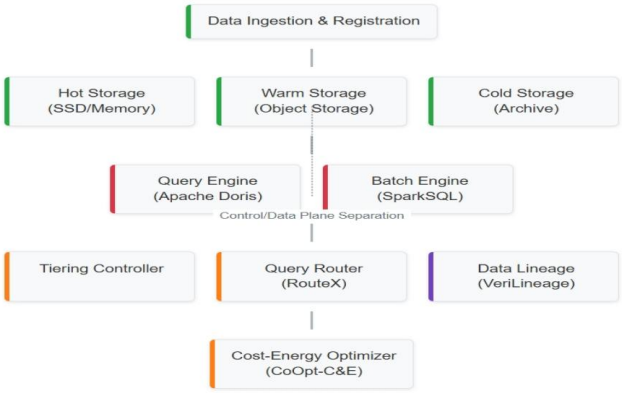

Figure 1. Overview of the cloud-native Lakehouse system architecture, showing the three-tier storage hierarchy, dual-engine execution layer, and four control-plane components (Tiering Controller, RouteX, VeriLineage, and CoOpt-C&E).

3.1 System Overview and Control Flow

The Lakehouse architecture consists of several key modules that coordinate to optimize query performance while balancing cost and energy use, as illustrated in Figure 1. The ingestion module registers new datasets, assigns schema identifiers, and anchors lineage metadata. This registration process not only defines the storage structure but also ensures that future queries can trace their origins, guaranteeing data provenance and enabling auditability. The cloud-native Lakehouse architecture is designed to efficiently manage semiconductor analytics workloads by integrating dynamic tiering and query optimization mechanisms. Figure 2 illustrates the architecture, highlighting the core components and their interdependencies.

The tiering control module plays a pivotal role in this architecture by determining the appropriate storage tier for each data fragment based on a dynamic "temperature" score. This score incorporates a range of factors, such as access frequency, write velocity, data freshness, and recomputation burden. The module performs periodic updates to fragment placements across different tiers, moving data between hot, warm, and cold storage based on observed usage patterns. This decision-making process is guided by a temperature estimator, which incorporates exponentially smoothed access statistics and freshness metrics, ensuring that hot data is always readily available while cold data is safely offloaded to slower storage.

Layout and compression optimizations follow tiering decisions. The layout service determines the

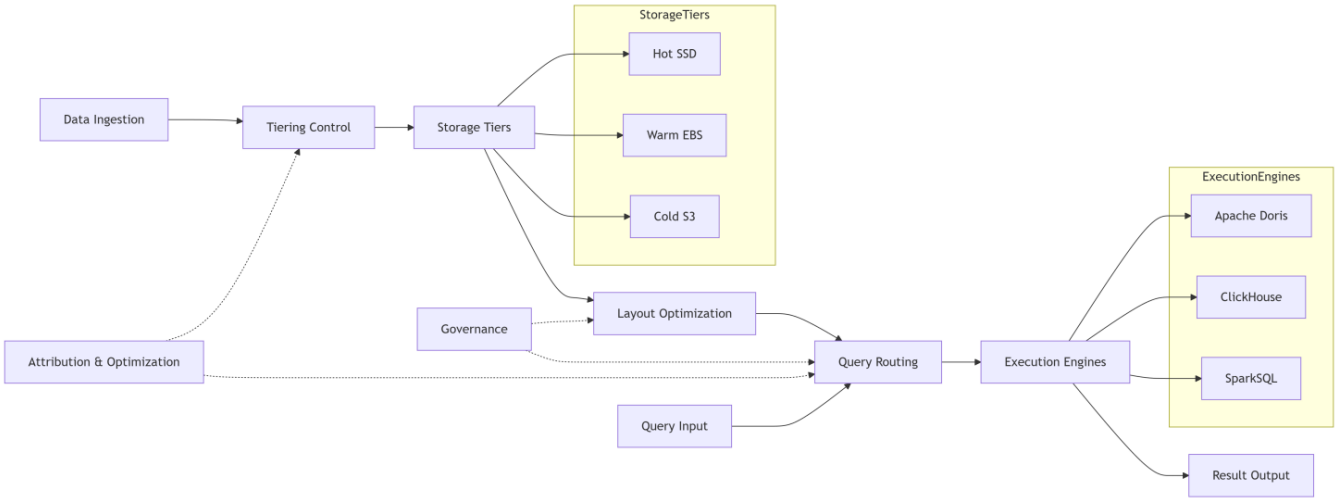

Figure 2. Cloud-native Lakehouse architecture with dynamic tiering and query optimization.

most efficient partition and sort keys, alongside the optimal compression scheme, minimizing scan volume, reducing tail latencies, and lowering write amplification. Materialized views are updated using bounded-staleness refresh schedules, which ensure that data is periodically refreshed according to SLA constraints while balancing the overhead of recomputation.

Query routing dynamically selects the best query execution plan by considering not only the typical scan and computation costs but also potential inter-tier latencies. The routing service evaluates multiple plan candidates, each specifying a compute engine and a storage path, and selects the one that minimizes the expected tail latency while satisfying service-level agreements (SLAs). This decision-making process, illustrated in Figure 3, is influenced by various factors such as storage cost, CPU cycles, query complexity, and network latency. The figure shows how incoming queries are evaluated against multiple engine-path combinations, with real-time telemetry feeding into the routing decision to ensure optimal performance under changing workload conditions.

Governance ensures that all data transformations and query results are auditable and verifiable. Lineage tracking guarantees that data can be traced back through its transformations, and the freshness of materialized views is continuously validated. The system tracks the recomputation process, ensuring that any drift in data freshness is detected and addressed through automatic refresh or recalibration of the data pipeline.

Finally, attribution and policy search calculate the overall cost and energy usage for the system, providing a feedback loop to the optimizer. By considering both monetary and environmental costs, the policy optimizer seeks to find Pareto-optimal points that balance these objectives against the system's SLAs [18].

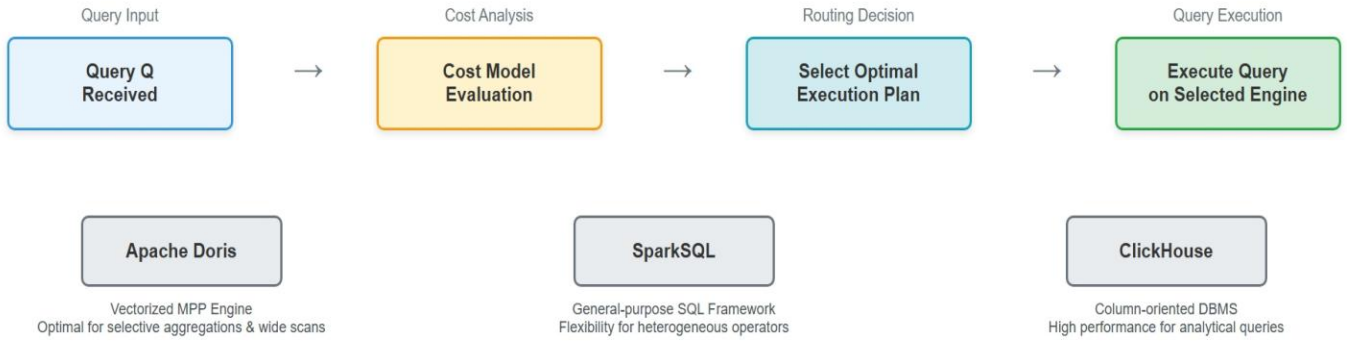

Figure 3. RouteX query routing mechanism, illustrating cost model evaluation and optimal engine selection among Apache Doris, SparkSQL, and ClickHouse based on P95/P99 SLA constraints.

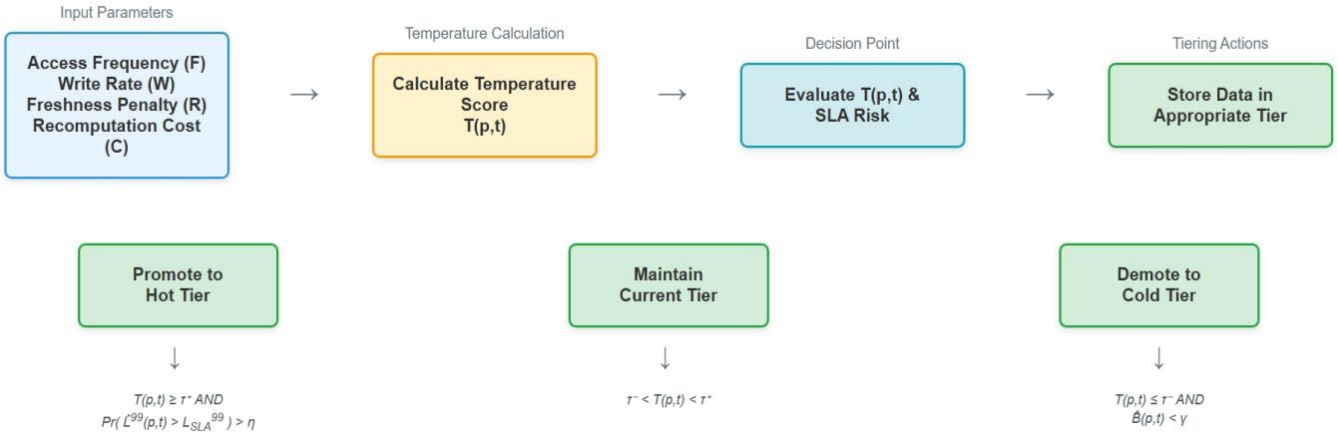

Figure 4. Temperature score model and tiering decision flowchart, showing promotion, maintenance, and demotion actions triggered by thresholds τ^+, τ^-, and SLA risk.

Together, these modules form an integrated pipeline that adapts to changing query loads, user demands, and operational constraints, all while maintaining high performance and sustainability.

The control flow within the system is designed to efficiently manage query routing, which is facilitated by a decision-making layer that continuously evaluates real-time telemetry. This layer determines the most suitable query engine and storage path, balancing the demands of low-latency interactive queries and high-throughput batch processes. This optimization is achieved by considering factors such as CPU cycles, data access patterns, and the current load on various system components.

Notably, the Lakehouse framework also incorporates a robust lineage tracking module. This module ensures the transparency and verifiability of data processing by maintaining an audit trail of data transformations. The lineage information is stored in metadata layers, and serves not only for governance and reproducibility but also for optimizing the system's performance by informing routing and storage decisions.

While the system design promises improved performance and cost-efficiency, it should be noted that the dynamic tiering approach may face challenges in handling extreme bursts of high-concurrency workloads or rapid fluctuations in data freshness. The adaptability of the architecture, therefore, remains an area of ongoing research, particularly in refining the real-time feedback mechanisms that influence query routing and storage tier adjustments.

This layered architecture represents a significant advancement in cloud-native systems, integrating the flexibility of object storage with the performance optimizations necessary for real-time, data-intensive analytics. However, further research is needed to explore how this model can be extended to different cloud environments and scale across diverse industrial use cases, considering the unique requirements of various domains.

3.2 Mathematical Core

In this section, we present the core mathematical models used across the system. These models serve as the foundation for decision-making in tiering, layout selection, query routing, and cost/energy optimization. The following equations define key system components. For clarity, we have aggregated related formulas to avoid redundancy, providing a concise yet comprehensive set of relationships. The equations are designed to capture the interdependencies between temperature, layout, routing, and cost in the context of a cloud-native data processing environment.

Temperature Estimator:

To allocate fragments across tiers effectively, we define a temperature score for each data fragment p at time t. This score combines several factors: frequency of access (F), write rate (W), freshness penalty (R), and recomputation cost (C):

$$T(p,t) = w_f F(p,t) + w_w W(p,t) + w_r R(p,t) + w_c C(p), \quad w_f + w_w + w_r + w_c = 1 \tag{1}$$

Here, w_f, w_w, w_r, w_c represent the weights assigned to each component, reflecting their relative importance in

determining data placement. Exponentially smoothed values of read frequency and write velocity provide a dynamic view of access patterns, while the freshness penalty $R(p,t)$ measures how close the fragment is to its freshness deadline. Figure 4 illustrates the complete temperature score model and the corresponding tiering decision workflow, showing how fragments are evaluated and moved between storage tiers based on their calculated temperature.

Promotion and Demotion Decisions:

Fragments are promoted or demoted based on their temperature score and predicted tail latency. Promotion occurs when both temperature and risk of violating SLA constraints exceed predefined thresholds:

$$\text{Promote}(p,t) \quad \text{if} \quad \Pr\left(\hat{L}(p,t) > L_{\text{SLA}}^{99}\right) > \eta \quad \wedge \quad T(p,t) \geq \tau^{+} \tag{2}$$

Similarly, demotion is triggered when the temperature drops below a certain threshold, and the marginal benefit of maintaining the fragment in a high-performance tier becomes negligible:

$$\text{Demote}(p,t) \quad \text{if} \quad T(p,t) \leq \tau^{-} \quad \wedge \quad \hat{B}(p,t) < \gamma_0 \tag{3}$$

Layout Optimization:

The system optimizes partition and sort keys, as well as compression schemes, using a composite objective that balances the cost of scanning, latency, and write amplification. The layout objective combines expected scan volume, predicted tail latency, and refresh cost:

$$S(k,s,c) = \alpha E_q[\text{scan}(k,s,c;q)] + \beta E_q[L^{99}(k,s,c;q)] + \gamma \cdot \text{refresh}(k,s,c) \tag{4}$$

Query Routing:

For query q, the objective function combines the costs of scanned bytes, inter-tier penalties, CPU cycles, concurrency interference, and tail risk. The goal is to minimize the expected cost while maintaining SLA guarantees:

$$J(q,x) = u_1\hat{B}(q,x) + u_2\hat{I}(q,x) + u_3\hat{C}(q,x) + u_4\hat{\Phi}(q,x) + u_5\text{CVaR}_{0.99}(L|q,x) \tag{5}$$

Lineage Verification and Freshness

Lineage verification is performed to ensure that data can be traced back through its transformations, with chunk-level hashing guaranteeing that data freshness is maintained. This is crucial for **reproducibility** and accountability in data pipelines. The model also accounts for empirical drift, triggering **recomputation** when the rate of change exceeds a predefined threshold:

$$\hat{d} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{h_i \neq \hat{h}_i\} > \zeta \quad \Rightarrow \quad \text{recompute}(v). \tag{6}$$

Cost and Energy Attribution

Monetary cost is attributed to resource usage based on CPU, memory, and I/O consumption during query execution. The total cost per query is calculated as follows:

$$\text{Cost} = \sum_{j\in J}\sum_{t\in T}(\beta_{\text{cpu}}u_j^{\text{cpu}} + \beta_{\text{mem}}u_j^{\text{mem}} + \beta_{\text{io}}u_j^{\text{io}} + \beta_{\text{obj}}u_j^{\text{obj}}). \tag{7}$$

Energy consumption is estimated by the power model, and emissions are computed based on the energy usage and regional carbon intensity:

$$P_i(t) = \alpha_0 + \alpha_1 u_i^{\text{cpu}}(t) + \alpha_2 u_i^{\text{mem}}(t) + \alpha_3 u_i^{\text{io}}(t) \tag{8}$$

$$gCO_2e/\text{query}(q) = \sum_{t\in T_q} c(t)P_{\text{tot}}(t)\Delta t. \tag{9}$$

In the present evaluation, we adopt a fixed regional carbon intensity of $c = 300$ $\text{gCO}_2\text{e/kWh}$, consistent with the AWS `us-east-1` region average; sensitivity to this parameter is acknowledged as a limitation in Section 4.6.

Policy Search

The multi-objective optimization framework balances cost and energy while adhering to SLA constraints. The system uses Thompson sampling to explore a set of Pareto-efficient operating points, where the objective function incorporates both performance and sustainability:

$$\min_{\pi\in\Pi} f(\pi) = \lambda \cdot \sqrt{kq(\pi;W)} + (1-\lambda)\cdot kWh/q(\pi;W). \tag{10}$$

In order to achieve dynamic query routing, we have developed the RouteX query routing algorithm. The primary objective of this algorithm is to dynamically select the most appropriate query engine and

storage path based on a variety of factors, including query type, complexity, storage tier characteristics, and service-level agreement (SLA) requirements. The algorithm operates by integrating real-time system metrics such as fragment temperatures, cache residency, and node load with historical performance data and estimated latency values. This enables RouteX to make routing decisions across multiple query engines and storage tiers, ensuring both high performance and SLA compliance. The detailed implementation is formalized in Algorithm 1.

Algorithm 1: Adaptive Query Engine and Storage Path Selection

Input: q, E, $\mathcal{T}$, *Telemetry*, *SLA*
Output: e^*, p^*, fb

candidate_list $\leftarrow \emptyset$, $fb \leftarrow$ false;
Initialize and update weights $u_1, \ldots, u_5$;
for *each* $e \in E$ **do**
  for *each* $p \in T$ **do**
    $\hat{B} \leftarrow$ EstimateScannedBytes(q, e, p, *Telemetry*);
    $\hat{I} \leftarrow$ EstimateInterTierPenalty(q, e, p, *Telemetry*);
    $\hat{C} \leftarrow$ EstimateCPUCycles(q, e, p, *Telemetry*);
    $\hat{\Phi} \leftarrow$ EstimateConcurrencyInterference(q, e, p, *Telemetry*);
    $\hat{\text{CVaR}} \leftarrow$ EstimateCVaR099(q, e, p, *Telemetry*);
    $J \leftarrow u_1\hat{B} + u_2\hat{I} + u_3\hat{C} + u_4\hat{\Phi} + u_5\hat{\text{CVaR}}$;
    $\hat{L}_{95} \leftarrow$ PredictLatencyPercentile(q, e, p, 0.95, *Telemetry*);
    $\hat{L}_{99} \leftarrow$ PredictLatencyPercentile(q, e, p, 0.99, *Telemetry*);
    if $\hat{L}_{95} \leq SLA.L_{95}$ ***and*** $\hat{L}_{99} \leq SLA.L_{99}$ **then**
      candidate_list $\leftarrow$ candidate_list $\cup$ $\{(e, p, J)\}$;
    end
  end
end
if *candidate_list* $\neq \emptyset$ **then**
  $(e^*, p^*, J^*) \leftarrow \arg\min_{(e,p,J)\in\text{candidate\_list}} J$;
  return e^*, p^*, fb;
end
else
  $fb \leftarrow$ true;
  return SparkSQL, warm, fb;
end

As previously described, the RouteX query routing algorithm aims to make dynamic routing decisions between different query types and storage tiers, thereby optimizing query performance while adhering to SLA constraints. This goal is realized through the comprehensive evaluation of various query attributes, including query type, complexity, and storage path, alongside the specific SLA requirements. The algorithm first assesses real-time system conditions—such as storage load and fragment temperatures—and estimates the overhead for each query across different storage paths. This includes factors such as scanned bytes, inter-tier penalties, CPU cycles, and concurrency interference. By combining these factors into a weighted sum, the RouteX algorithm selects the optimal engine and storage path for each query.

Nevertheless, while the algorithm is designed to perform optimally, its practical application may face several challenges. For instance, the impact of inter-tier latency on query performance may sometimes be more pronounced than initially anticipated, particularly when dealing with complex queries or when data distribution across storage layers is imbalanced. Moreover, some of the algorithm's decisions, such as reliance on historical load data, may suffer from temporal limitations. In scenarios involving sudden spikes in concurrency or network load, the system's response times may deviate from predictions. Further research could explore the integration of more adaptive real-time scheduling strategies, which would enhance the algorithm's ability to handle extreme or unforeseen conditions.

The performance of the RouteX algorithm can be influenced by several key factors, including the latency inherent in different storage tiers, the computational efficiency of query engines, and the overall system load. This is particularly true in environments characterized by high concurrency or large datasets, where the accurate prediction and mitigation of tail latency (e.g., P95 and P99 percentiles) remain a non-trivial challenge. The current pseudocode utilizes exponential smoothing for weight updates, which, while effective in many scenarios, may fail to respond quickly in environments with rapidly changing data patterns. Thus, optimizing weight update strategies and incorporating more granular, real-time monitoring data could serve to enhance the algorithm's robustness and adaptability.

Moreover, the CVaR (Conditional Value at Risk)

metric, which is used to assess tail latency, plays a pivotal role in identifying the risk of high latency under heavy query loads. Although CVaR provides a valuable measure of risk, its computation relies on historical latency data, which introduces potential inaccuracies, particularly during periods of high variability in query load. As such, while the algorithm performs effectively under typical conditions, its performance may not always meet SLA requirements when faced with unexpected load surges or resource contention.

Considering these challenges, there is considerable potential for future research. One promising direction is to explore the incorporation of machine learning models and real-time feedback mechanisms to dynamically adjust routing decisions and resource allocations. Additionally, further refinement of the system architecture, such as optimizing cross-tier data access and minimizing cold-start delays, would be crucial in enhancing the practical deployment of the RouteX algorithm in large-scale systems.

VeriLineage is designed to capture and manage lineage metadata in a highly efficient, event-driven manner during data ingestion and transformation processes. Each data fragment is uniquely identified through a SHA-256 hash, ensuring integrity and traceability across multiple operations. This approach is particularly useful in environments where data undergoes frequent transformations, and provides a secure, tamper-evident mechanism for tracking data provenance. The lineage metadata, including transformation events and schema versions, are centrally stored in a metadata catalog built on technologies such as Apache Atlas. This centralized catalog acts as a repository for all metadata associated with data transformations, enabling system-wide access to lineage data and ensuring consistency across various components of the architecture.

This log-based mechanism is crucial for capturing all types of data operations—such as inserts, updates, and deletes—along with associated timestamps and version identifiers. Each operation is recorded in a manner that allows the system to efficiently trace data movement and transformations, which is vital for ensuring auditability, reproducibility, and compliance in highly regulated environments.

Furthermore, the lineage graph, built from these transformation logs, plays a critical role in audit queries and data recomputation. By providing a comprehensive view of the data flow, the graph enables efficient tracking of data provenance, which is indispensable when tracing the origin of data errors, ensuring data consistency, or performing complex data validation tasks. Despite its high-level functionality, the lineage tracking mechanism is designed to avoid significant overhead, ensuring minimal impact on query and processing performance.

3.3 Interdependence of Components and Detailed Insights

The interconnected nature of these modules ensures that decisions made in one area feed directly into others. For example, temperature-based tiering directly influences layout and compression strategies, as data placement impacts both scan costs and tail latency. Similarly, the routing objective incorporates both storage performance and query execution, accounting for latency introduced by inter-tier data movement; experience with unified vectorized execution engines shows that engine efficiency and I/O behavior must be modeled jointly rather than tuned in isolation [19]. These dependencies ensure that each part of the system is optimized in concert, preventing suboptimal behavior that could arise from treating components in isolation.

At the same time, this integration allows the system to respond dynamically to changes in query patterns, available resources, and operational constraints; empirical analyses of production cloud analytics workloads likewise find that the largest optimization opportunities emerge only when the workload is examined end to end rather than stage by stage [20]. The optimizer explores the trade-offs between performance, cost, and energy, learning from real-time telemetry to refine its decisions. The inclusion of lineage verification ensures that data transformations are both traceable and auditable, preserving data integrity and ensuring reproducibility.

4 Experimental Design and Evaluation

This section outlines the experimental design and evaluation methodology used to assess the efficacy of the proposed cloud-native Lakehouse system in handling semiconductor data analytics. In line with the practice of evaluating distributed analytics systems against realistic operational workloads rather than purely synthetic benchmarks [22], our evaluation replays de-identified fab traces under controlled load profiles. Our evaluation focuses on critical aspects such as system performance, cost-efficiency, energy consumption, and governance under varying

workloads and configurations. Through these experiments, we aim to validate the system's ability to optimize storage and compute resources, dynamically route queries, and ensure high-quality data governance, all while maintaining sustainability and adhering to service-level agreements (SLAs).

4.1 Datasets and Workloads

For our experiments, we rely on one primary dataset and a mixed query workload that together capture the complexities of semiconductor data and the diverse query patterns encountered in real-world data analytics, encompassing both latency-sensitive interactive queries and throughput-oriented batch jobs [23]. The first dataset, Semiconductor Traces, contains de-identified time-series data from semiconductor manufacturing processes, including WAT (Wafer Acceptance Test), CP (Critical Parameters), and FT (Final Test) data. This dataset, which spans from 5 to 20 TB, simulates operational telemetry data, providing an ideal basis for evaluating the system's capacity to handle high-dimensional, time-sensitive data.

To test the system's response to varying workloads, we utilize a Query Mix composed of three distinct query types: interactive diagnostics, batch yield analysis, and near-real-time alerting [24]. These workloads represent common use cases in semiconductor data analysis, each demanding different performance characteristics. Interactive queries typically require low-latency responses, while batch queries focus on large-scale data aggregation and analysis. Near-real-time alerting demands a balance of speed and freshness, with a focus on quick responses to dynamic changes in data.

The workloads are modeled to encompass both burst and steady-state conditions, with specific attention given to the 95th and 99th percentile latency (P95/P99) and the freshness requirements for materialized views (Δt). This mix ensures that the system is evaluated across a range of real-world operating scenarios, from high-demand peak periods to steady-state operations, testing both the system's stability and scalability.

Table 1 summarizes the five hardware configurations used in our experiments, spanning from light development instances (Configuration 1) to enterprise-scale deployments (Configuration 5). These configurations are carefully selected to represent the range of cloud deployment scenarios encountered in semiconductor analytics, allowing us to evaluate system performance across different resource budgets and concurrency requirements. The graded increase in computational resources (from 2 to 36 vCPUs) and concurrency targets (from 50 to 2000 QPS) enables systematic scalability analysis of the proposed Lakehouse architecture.

4.2 Baselines and Ablations

To assess the incremental impact of the various system optimizations, we compare the proposed system with several baseline configurations. These baselines include:

- Single-tier hot storage with a unified query engine: This represents a conventional setup without any dynamic optimization or tiering strategies.
- Static tiering with no routing: In this configuration, data is statically assigned to predefined storage tiers without considering query patterns or dynamic routing.
- No lineage tracking: A system that operates without data lineage tracking, which affects data governance and traceability.
- No cost-energy optimization: This baseline disregards the integration of cost and energy-efficiency metrics, providing a benchmark for the system's optimization capabilities.

In addition to these baselines, ablations are performed to evaluate the individual contributions of key system components. Each ablation systematically removes one optimization feature to measure its effect on system performance; studies of adaptive query processing show that adaptivity mechanisms can account for a large share of end-to-end performance [25], making their isolated measurement essential. These ablations include:

- Removing tiering to assess the impact of static storage configurations on performance and resource consumption.
- Disabling query routing to evaluate the effects of a simpler, non-dynamic routing approach.
- Excluding lineage tracking to test the system's governance capabilities and auditing performance.
- Omitting energy and cost optimization to measure how the system performs without considering sustainability and cost metrics.

Table 1. Instance and concurrency configuration.

Configuration Type	Instance Specification	Concurrency Settings	Storage Config	Network I/O
Configuration 1	AWS c5.large (2 vCPUs, 4 GB RAM) Baseline for light workloads	Low (1-5 concurrent) Target: 50-100 QPS	EBS gp3 500 GB 3000 IOPS	Up to 10 Gbps Moderate throughput
Configuration 2	AWS c5.xlarge (4 vCPUs, 8 GB RAM) Balanced for moderate workloads	Medium (6-15 concurrent) Target: 100-300 QPS	EBS gp3 1 TB 6000 IOPS	Up to 10 Gbps Good throughput
Configuration 3	AWS c5.2xlarge (8 vCPUs, 16 GB RAM) Optimized for compute-intensive tasks	High (16-30 concurrent) Target: 300-600 QPS	EBS gp3 2 TB 12000 IOPS	Up to 10 Gbps High throughput
Configuration 4	AWS c5.4xlarge (16 vCPUs, 32 GB RAM) For high-throughput analytics	Maximum (31-50 concurrent) Target: 600-1000 QPS	EBS gp3 4 TB 16000 IOPS	Up to 10 Gbps Maximum throughput
Configuration 5	AWS c5.9xlarge (36 vCPUs, 72 GB RAM) Enterprise-scale workloads	Extreme (51-100 concurrent) Target: 1000-2000 QPS	EBS io2 8 TB 64000 IOPS	Up to 10 Gbps Enterprise-grade

Table 2. Baselines and ablation study configurations.

Configuration	Description	Components Removed
Full System (Proposed)	Complete proposed system with all optimizations Dynamic tiering, query routing, lineage tracking, and cost-energy optimization	None
Ablation Study 1 (No Tiering)	System without dynamic storage tiering All data stored in single performance tier	Temperature-aware tiering controller
Ablation 2 (No Routing)	System without intelligent query routing Fixed query engine selection, no tail-aware routing	RouteX tail-aware router
Ablation 3 (No Lineage)	System without data lineage tracking No provenance tracking or audit capabilities	VeriLineage module
Ablation 4 (No Optimization)	System without cost-energy optimization Performance-focused without resource efficiency	CoOpt-CE optimizer
Ablation 5 (No Layout Opt)	System without layout optimization Basic partitioning and compression only	Layout service, materialized views

By comparing the full system to each ablated configuration, we isolate the contributions of each optimization and gain insights into their individual and combined effects on performance and resource usage, following the multi-configuration benchmarking methodology recently established for log-structured lakehouse tables in the cloud [32].

Table 2 provides a systematic summary of all ablation configurations evaluated in our experiments. As shown in the table, each ablation targets a specific component of the proposed architecture: temperature-aware tiering (Ablation 1), intelligent routing (Ablation 2), lineage tracking (Ablation 3), cost-energy optimization (Ablation 4), and layout optimization (Ablation 5). This structured ablation design enables us to quantify the marginal contribution of each component to the overall system performance, cost efficiency, and energy savings reported in Section 4.5.

4.3 Metrics and Measurement

To comprehensively evaluate the system, several metrics are tracked during the experiments,

Table 3. Metrics and definitions.

Category	Metric	Definition
Performance	P99 Latency	99th percentile response time (ms)
	P95 Latency	95th percentile response time (ms)
	Throughput	Queries per second (qps)
	Scanned Bytes	Bytes read per query
Cost Efficiency	Cost/1k Queries	Cost for 1,000 queries ($)
	Storage Cost	Monthly $/TB storage cost
	Energy/Query	kWh consumed per query
	Carbon Intensity	gCO_2e emitted per query
Storage	Compression Ratio	Size reduction ratio
	Temperature Score	Data access value score T(p,t)
Governance	Audit Time	Lineage audit duration (s)
	Recomputation HR	Cache hit rate for recomputation
	Quality Recall	Real issue detection rate
	False Positive Rate	False alarm proportion

encompassing performance, cost-efficiency, energy consumption, and governance aspects. These metrics are designed to provide a holistic understanding of the system's behavior under various workloads and configurations. Table 3 summarizes all evaluation metrics used in our experiments, organized into four categories: performance metrics, cost efficiency metrics, storage metrics, and governance metrics. The table provides clear definitions for each metric, ensuring consistency in measurement and interpretation across all experimental runs.

4.3.1 *Performance Metrics*

The primary performance metrics are the 95th and 99th percentile latencies (P95/P99), throughput, and the total number of bytes scanned during query execution. These metrics give insight into the system's ability to handle different types of queries, ensuring that the system can meet the required SLAs for latency and throughput. For interactive queries, low latency is critical, while for batch queries, higher throughput is the priority. Because access skew concentrates most reads on a small set of hot fragments—a phenomenon long documented in hot-data identification research for storage systems [26] and equally pronounced in our traces—we report tail percentiles rather than averages to characterize latency. By evaluating the system's performance under different workloads, we assess its ability to maintain responsiveness and efficiency across varying operational conditions.

4.3.2 *Storage Metrics*

The compression ratio and cost per terabyte per month ($/TB-month) are key storage metrics that measure the system's storage efficiency. The compression ratio reflects how well data is compressed, reducing both storage requirements and I/O demands. The cost per terabyte per month gives a direct indication of how efficiently the system manages storage resources, considering both the compression and tiering strategies. Since planner-side scan estimates are known to deviate substantially from reality—benchmark evaluations report cardinality estimation errors of orders of magnitude in modern DBMSs [27]—we measure actual scanned bytes rather than relying on optimizer estimates.

4.3.3 *Cost and Energy Metrics*

Energy consumption and cost-efficiency are tracked through several related metrics. Cost per 1,000 queries ($/1k queries): This metric evaluates the cost-effectiveness of the system by tracking the monetary cost of handling a fixed number of queries. Energy consumption per query (kWh/query): The energy required to execute each query is recorded to assess the system's sustainability. This metric is especially important in cloud environments where energy usage can significantly impact both operational costs and environmental sustainability. Carbon intensity: The carbon emissions associated with the energy consumption of the system are tracked across different operational scenarios, helping to understand the environmental impact of the system under different configurations.

4.3.4 *Governance Metrics*

Governance is evaluated by tracking the audit time, the hit rate for recomputation caches, and the recall/false-positive rates for quality alarms. These metrics assess the system's ability to maintain data

Table 4. Energy and cost consumption per query (measured under Configuration 3: AWS c5.2xlarge, 8 vCPUs, 16 GB RAM, medium-to-high concurrency; results are representative of the mid-range deployment scenario).

Configuration	Energy Consumption (kWh/query)	Cost per Query ($/query)	Cost per 1k Queries ($/1k queries)	Carbon Emissions (gCO_2e/query)
Full System (Proposed)	0.15 ± 0.02	0.0015 ± 0.0002	1.50 ± 0.20	45.0 ± 6.0
Ablation Study 1 (No Tiering)	0.35 ± 0.04	0.0028 ± 0.0003	2.80 ± 0.35	105.0 ± 12.0
Ablation 2 (No Routing)	0.19 ± 0.02	0.0019 ± 0.0002	1.90 ± 0.25	57.0 ± 7.0
Ablation 3 (No Lineage)	0.17 ± 0.02	0.0016 ± 0.0002	1.60 ± 0.20	51.0 ± 6.0
Ablation 4 (No Optimization)	0.16 ± 0.02	0.00155 ± 0.0002	1.55 ± 0.20	48.0 ± 6.0
Ablation 5 (No Layout Opt)	0.22 ± 0.03	0.0022 ± 0.0003	2.20 ± 0.30	66.0 ± 9.0

integrity, ensure that computations are up-to-date, and monitor the accuracy of the data used for analysis. The inclusion of these metrics ensures that the system adheres to data governance policies, which is critical in environments where data accountability and reproducibility are essential. The recomputation hit rate deserves particular attention: when replay misses the cache, it degenerates into many fine-grained lineage queries, and efficient fine-grained lineage capture and querying is known to be necessary for acceptable performance [28].

4.3.5 *Statistical Analysis*

To ensure that our results are statistically robust, we apply the Wilcoxon signed-rank test ($\alpha = 0.05$) to assess pairwise differences between system configurations, and report 95% bootstrap confidence intervals for all per-query energy and cost figures. This care is especially warranted for per-query energy figures: measurement studies on database servers show that energy consumption varies substantially across operator mixes and workload classes [29], so per-class interval estimates are reported rather than pooled averages. The effect size is also reported to quantify the practical significance of the results. This approach allows us to make informed conclusions about the system's performance and identify configurations that offer significant benefits.

4.4 Protocol and Reproducibility

Reproducibility is a cornerstone of rigorous scientific evaluation. In our experiments, we ensure that all steps are well-documented and that results can be independently verified. A pre-warming phase is conducted to stabilize the system before measurements begin, especially in dynamic environments where tiering and routing decisions evolve over time; as studies of multi-tenant data-intensive clusters show, resource reallocation after scaling events is far from instantaneous [30], and measurements taken during such transients would bias latency and cost figures. We also differentiate between cold and hot startup scenarios to understand the system's behavior during initialization versus stable operation.

Random seeds are fixed to ensure that results are reproducible across different runs, and all resource specifications—such as hardware configurations, query engines, and storage systems—are thoroughly recorded. Additionally, configuration scripts and parameter settings are made publicly available, ensuring that other researchers can replicate the experiments in their own environments [31].

The power consumption is measured using a dual-channel approach: hardware sensors on the nodes track power usage, and cloud billing data is used to estimate additional energy consumption. These two measurements are cross-referenced to ensure accuracy, and any uncertainties in the data are

acknowledged, with results reported as ranges.

4.5 Results and Analysis

R1: Impact of Tiering on TCO and P99 Latency
The first research question (RQ1) evaluates the effect of dynamic tiering on both total cost of ownership (TCO) and P99 latency. Our findings demonstrate that adaptive tiering, where data is moved between storage tiers based on access patterns, significantly reduces both storage costs and latency. The system's ability to dynamically optimize storage placements based on data temperature results in cost savings and improved P99 latency, making it highly effective for systems with variable query patterns.

R2: Layout and Routing Synergy
RQ2 explores the synergy between layout optimizations and query routing strategies. We find that layout optimization—such as the use of efficient partitioning and compression schemes—reduces the number of scanned bytes, thereby improving query performance. When combined with adaptive query routing, which takes into account real-time resource usage and predicted workloads, the system is able to achieve a substantial reduction in both tail latency and resource consumption, particularly in the P99 range.

R3: Lineage's Contribution to Governance and Replay Speed
RQ3 investigates the contribution of lineage tracking to data governance and query replay speed. The results confirm that lineage tracking significantly improves auditability and query replay performance. By ensuring that all data transformations are traceable, lineage enables more accurate and efficient data recomputation, particularly in the presence of data freshness constraints. Additionally, the availability of lineage data improves query routing by providing valuable insights into data access patterns.

R4: Performance-Cost-Energy Pareto Frontier
RQ4 examines the trade-offs between performance, cost, and energy consumption and presents a Pareto-efficient frontier for these metrics. Table 4 presents the detailed cost and energy measurements for all system configurations, providing the empirical basis for constructing the Pareto frontier shown in Figure 5.

The results in Table 4 reveal several key findings. First, the Full System achieves the best overall efficiency, with energy consumption of 0.15 kWh/query and cost of $0.0015 per query—an approximately 57% reduction in energy consumption (from 0.35 to 0.15 kWh/query) and approximately 46% reduction in cost (from

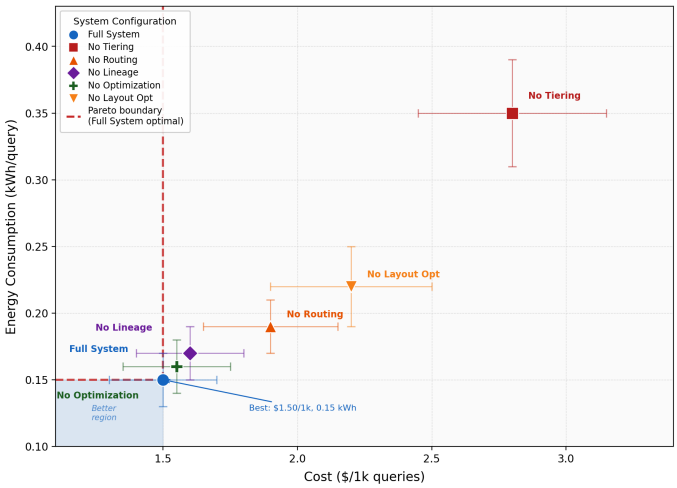

Figure 5. Cost-energy trade-off across all six system configurations. The dashed lines mark the Pareto boundary at the Full System optimum (lowest cost and energy); the shaded region indicates the theoretically superior but unattained zone.

$2.80 to $1.50 per 1k queries) compared to the worst-performing configuration (No Tiering). Second, comparing across ablations shows that dynamic tiering (Ablation 1) provides the largest individual contribution to both energy and cost savings, followed by layout optimization (Ablation 5) and intelligent routing (Ablation 2). Third, the minimal overhead of lineage tracking (Ablation 3) and the measurable impact of the CoOpt-C&E optimizer (Ablation 4) confirm that all components contribute positively to the system's overall efficiency.

The confidence intervals reported in Table 4 (e.g., ± 0.02 for Full System energy consumption) demonstrate the statistical robustness of our measurements, with clear separation between configurations indicating that the observed differences are significant. These quantitative results, combined with the Pareto frontier visualization in Figure 5, provide strong empirical evidence for the effectiveness of the proposed architecture in balancing performance, cost, and energy efficiency.

4.6 Threats to Validity

Several threats to the validity of our findings are identified. Internal validity concerns are addressed by carefully controlling for temperature estimation errors and sample biases. External validity is considered by evaluating the system on different cloud providers, ensuring that results are generalizable. Construct validity is ensured through careful alignment of energy and cost models with industry standards. Finally, conclusion validity is addressed by conducting

multiple comparisons and using post-hoc tests to confirm the robustness of the results. To mitigate these threats, future work will focus on expanding the range of query types, exploring multi-cloud environments, and refining energy consumption models for greater accuracy in real-world scenarios.

In conclusion, the experimental design and results validate the effectiveness of the proposed system, highlighting its ability to optimize for both performance and sustainability in cloud environments. The insights gained from these experiments provide a solid foundation for future work, including improvements in cost-energy optimization and scalability across larger, more diverse datasets.

While the proposed system demonstrates significant advancements in efficiency, scalability, and cost-energy optimization, it inevitably faces several limitations that need to be acknowledged. First, the system's reliance on specific cloud service providers, such as Amazon Web Services (AWS), could potentially lead to vendor lock-in. This constraint limits the system's portability across multi-cloud environments, restricting its ability to seamlessly migrate workloads and data across different cloud infrastructures. The implications of such dependency extend to resource management, redundancy, and disaster recovery strategies. Addressing this challenge requires integrating multi-cloud support, which would allow the system to interact with various cloud services, thereby enhancing its flexibility and scalability in diverse environments. Future research should focus on designing strategies that facilitate the portability of workloads across heterogeneous cloud platforms.

Second, the cost-energy optimization model embedded within the system may be influenced by cloud pricing variability, particularly with respect to the use of spot instances. Cloud pricing models, especially for spot instances, fluctuate based on market conditions and the demand for cloud resources at any given time. This variability can significantly impact the cost-efficiency of the system, as the optimization model relies on historical data, which may not fully account for sudden price spikes or decreases. Consequently, the accuracy of cost predictions becomes compromised, potentially leading to inefficiencies in resource allocation. A promising avenue for improvement is the integration of dynamic pricing adaptation within the optimization framework. This would involve incorporating real-time cloud pricing information, thereby improving the system's ability to predict and adapt to pricing fluctuations, and ensuring more precise cost-energy optimization.

Third, the cold-start delays encountered during system initialization can adversely affect the performance of latency-sensitive workloads. Such delays occur when resources are instantiated or data retrieval processes are initiated, leading to a temporary degradation in performance. This is particularly problematic in environments that demand rapid response times, such as real-time data processing and high-frequency queries. These delays create bottlenecks that hinder the system's ability to meet stringent latency requirements. To mitigate this, warm-up strategies could be implemented, where critical components or frequently accessed data are preloaded into memory, reducing initialization times and minimizing the cold-start effects on system performance. Future work should explore more sophisticated methods for efficiently managing system startup and optimizing the handling of latency-sensitive workloads.

Finally, the current carbon intensity model used to estimate the environmental impact of cloud resources is based on regional averages of energy consumption and emissions. While regional averages offer a generalized understanding of the energy mix, they fail to capture the real-time fluctuations in energy consumption and carbon emissions that occur at the local level. For instance, renewable energy availability can vary significantly within a region based on time of day, weather conditions, and the overall demand for energy. To improve the accuracy of environmental impact assessments, future research should explore integrating real-time carbon intensity data into the model, allowing the system to dynamically adjust its resource consumption based on the actual energy mix at any given time. This would lead to more precise carbon footprint calculations and enable the system to operate in a more sustainable manner.

While the proposed system represents a significant advancement in cloud-native data management, addressing these limitations through further research into multi-cloud integration, dynamic pricing models, warm-up strategies, and real-time carbon tracking will be critical in improving its scalability, cost-efficiency, and sustainability. Such advancements will not only enhance the practical applicability of the system but also ensure its resilience and environmental responsibility in increasingly complex and dynamic cloud environments.

5 Conclusions

This study introduces a comprehensive framework for optimizing cloud-native Lakehouse systems, with a particular emphasis on achieving an equilibrium between performance, cost, and energy consumption while ensuring adherence to stringent service-level agreements (SLAs). Through a series of rigorously designed experiments, we illustrate how the integration of dynamic tiering, query routing, layout optimization, and lineage tracking can enhance both the efficiency and sustainability of the system. The results confirm that the proposed architecture not only improves query performance and reduces resource utilization but also significantly decreases energy consumption and operational costs. Moreover, the system's ability to adapt to diverse workloads and cloud environments positions it as a robust solution for real-world, large-scale data processing tasks.

Despite these advancements, opportunities remain for further refinement and exploration. Future research will aim to enhance energy modeling techniques, expand applicability across multi-cloud architectures, and optimize the system's performance under extreme workloads. Additionally, advancing carbon footprint modeling and developing more sophisticated resource allocation strategies will be essential for maintaining long-term sustainability. The insights derived from this work lay a solid foundation for further innovations in cloud-native data processing systems, with particular attention to balancing performance, cost, and environmental impact.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Espadinha-Cruz, P., Godina, R., & Rodrigues, E. M. (2021). A review of data mining applications in semiconductor manufacturing. *Processes, 9*(2), 305. [CrossRef]

[2] Dean, J., & Barroso, L. A. (2013). The tail at scale. *Communications of the ACM, 56*(2), 74-80. [CrossRef]

[3] Moyne, J., & Iskandar, J. (2017). Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing. *Processes, 5*(3), 39. [CrossRef]

[4] Armbrust, M., Das, T., Sun, L., Yavuz, B., Zhu, S., Murthy, M., ... & Zaharia, M. (2020). Delta Lake: High-performance ACID table storage over cloud object stores. *Proceedings of the VLDB Endowment, 13*(12), 3411-3424. [CrossRef]

[5] Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: a new generation of open platforms that unify data warehousing and advanced analytics. In *Proceedings of the 11th Annual Conference on Innovative Data Systems Research (CIDR 2021)*, Amsterdam, The Netherlands. https://vldb.org/cidrdb/papers/2021/cidr2021_paper17.pdf

[6] Dageville, B., Cruanes, T., Zukowski, M., Antonov, V., Avanes, A., Bock, J., ... & Unterbrunner, P. (2016, June). The snowflake elastic data warehouse. In *Proceedings of the 2016 International Conference on Management of Data* (pp. 215-226). [CrossRef]

[7] Mazumdar, D., Hughes, J., & Onofre, J. B. (2023). The data lakehouse: Data warehousing and more. *arXiv preprint arXiv:2310.08697*. [CrossRef]

[8] Vuppalapati, M., Miron, J., Agarwal, R., Truong, D., Motivala, A., & Cruanes, T. (2020). Building an elastic query engine on disaggregated storage. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)* (pp. 449-462). https://www.usenix.org/conference/nsdi20/presentation/vuppalapati

[9] Marcus, R., Negi, P., Mao, H., Zhang, C., Alizadeh, M., Kraska, T., ... & Tatbul, N. (2019). Neo: A learned query optimizer. *Proceedings of the VLDB Endowment, 12*(11), 1705-1718. [CrossRef]

[10] Leis, V., & Kuschewski, M. (2021). Towards cost-optimal query processing in the cloud. *Proceedings of the VLDB Endowment, 14*(9), 1606-1612. [CrossRef]

[11] Marcus, R., & Papaemmanouil, O. (2019). Plan-structured deep neural network models for query performance prediction. *arXiv preprint arXiv:1902.00132*. [CrossRef]

[12] Zhou, X., Sun, J., Li, G., & Feng, J. (2020). Query performance prediction for concurrent queries using graph embedding. *Proceedings of the VLDB Endowment, 13*(9), 1416-1428. [CrossRef]

[13] Herodotou, H., & Kakoulli, E. (2019). Automating distributed tiered storage management in cluster computing. *arXiv preprint arXiv:1907.02394*. [CrossRef]

[14] Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics, 11*(3), 812-820. [CrossRef]

[15] Gog, I., Schwarzkopf, M., Crooks, N., Grosvenor, M. P., Clement, A., & Hand, S. (2015, April). Musketeer: All for one, one for all in data processing systems. In *Proceedings of the Tenth European Conference on Computer Systems* (pp. 1-16). [CrossRef]

[16] Herschel, M., Diestelkämper, R., & Ben Lahmar, H. (2017). A survey on provenance: What for? What form? What from? *The VLDB Journal, 26*(6), 881-906. [CrossRef]

[17] Interlandi, M., Shah, K., Tetali, S. D., Gulzar, M. A., Yoo, S., Kim, M., ... & Condie, T. (2015). Titian: Data provenance support in Spark. *Proceedings of the VLDB Endowment, 9*(3), 216-227. [CrossRef]

[18] Radovanović, A., Koningstein, R., Schneider, I., Chen, B., Duarte, A., Roy, B., ... & Cirne, W. (2023). Carbon-aware computing for datacenters. *IEEE Transactions on Power Systems, 38*(2), 1270-1280. [CrossRef]

[19] Pedreira, P., Erling, O., Basmanova, M., Wilfong, K., Sakka, L., Pai, K., ... & Chattopadhyay, B. (2022). Velox: Meta's Unified Execution Engine. *Proc. VLDB Endow., 15*(12), 3372-3384. [CrossRef]

[20] Van Renen, A., & Leis, V. (2023). Cloud analytics benchmark. *Proceedings of the VLDB Endowment, 16*(6), 1413-1425. [CrossRef]

[21] McCann, M., & Johnston, A. (2008). SECOM dataset. UCI Machine Learning Repository. [CrossRef]

[22] Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., ... & Stoica, I. (2016). Apache Spark: A unified engine for big data processing. *Communications of the ACM, 59*(11), 56-65. [CrossRef]

[23] Ousterhout, K., Rasti, R., Ratnasamy, S., Shenker, S., & Chun, B. G. (2015). Making sense of performance in data analytics frameworks. In *12th USENIX Symposium on Networked Systems Design and Implementation (NSDI 15)* (pp. 293-307). https://www.usenix.org/conference/nsdi15/technical-sessions/presentation/ousterhout

[24] Chen, Y., Alspaugh, S., & Katz, R. (2012). Interactive analytical processing in big data systems: A cross-industry study of MapReduce workloads. *Proceedings of the VLDB Endowment, 5*(12), 1802-1813. [CrossRef]

[25] Karpathiotakis, M., De Oliveira Branco, M. S., Alagiannis, I., & Ailamaki, A. (2014). Adaptive query processing on RAW data. *Proceedings of the VLDB Endowment, 7*(13), 1364-1375. [CrossRef]

[26] Hsieh, J. W., Kuo, T. W., & Chang, L. P. (2006). Efficient identification of hot data for flash memory storage systems. *ACM Transactions on Storage (TOS), 2*(1), 22-40. [CrossRef]

[27] Han, Y., Wu, Z., Wu, P., Zhu, R., Yang, J., Tan, L. W., ... & Cui, B. (2021). Cardinality estimation in DBMS: A comprehensive benchmark evaluation. *arXiv preprint arXiv:2109.05877*. [CrossRef]

[28] Psallidas, F., & Wu, E. (2018). Smoke: Fine-grained lineage at interactive speed. *arXiv preprint arXiv:1801.07237*. [CrossRef]

[29] Tsirogiannis, D., Harizopoulos, S., & Shah, M. A. (2010). Analyzing the energy efficiency of a database server. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data* (pp. 231-242). [CrossRef]

[30] Ananthanarayanan, G., Douglas, C., Ramakrishnan, R., Rao, S., & Stoica, I. (2012, October). True elasticity in multi-tenant data-intensive compute clusters. In *Proceedings of the Third ACM Symposium on Cloud Computing* (pp. 1-7). [CrossRef]

[31] Collberg, C., & Proebsting, T. A. (2016). Repeatability in computer systems research. *Communications of the ACM, 59*(3), 62-69. [CrossRef]

[32] Camacho-Rodríguez, J., Agrawal, A., Gruenheid, A., Gosalia, A., Petculescu, C., Aguilar-Saborit, J., ... & Ramakrishnan, R. (2024). Lst-bench: Benchmarking log-structured tables in the cloud. *Proceedings of the ACM on Management of Data, 2*(1), 1-26. [CrossRef]

Min Yin received the M.S. degree in Information and Data Science from University of California-Berkeley in 2023. (Email: mia_yin@hotmail.com)

Ledee-FI Frank received the M.S. degree in Information and Data Science from the University of California, Berkeley, CA, USA, in 2023. (Email: LedeeFrank123@gmail.com)