ISSN: 2096-3246
Volume 58, Issue 02, March, 2026

GREEN COMPUTING AND CARBON FOOTPRINT REDUCTION: A COMPREHENSIVE FRAMEWORK FOR SUSTAINABLE IT INFRASTRUCTURE MANAGEMENT

Krishan Singh[1], Touseef Ahmad Lone[2], Arvin Vinayek[3]
[1]Research Scholar, Department of Computer Science and Engineering, CT University, Ludhiana 142024, India
[2]Assistant Professor, Department of Computer Science and Engineering, CT University, Ludhiana 142024, India
[3]Assistant Professor, Department of Computer Science and Engineering, CT University, Ludhiana 142024, India

Email: writetokrish.97@gmail.com, Lonetouseef99@gmail.com, avinayek123@gmail.com

Abstract— The explosive proliferation of digital services, cloud computing platforms, and artificial intelligence workloads has transformed the information and communication technology (ICT) sector into one of the most energy-intensive industries on the planet. Global data centers consumed an estimated 415 terawatt-hours of electricity in 2024, and projections indicate this figure will surpass 900 terawatt-hours by 2030 if present growth trajectories continue unchecked. Concurrently, the embodied carbon locked into semiconductor manufacturing—a dimension frequently neglected in sustainability reporting—can exceed the entire operational carbon of a computing device over its useful lifetime. This paper presents the Integrated Carbon Management Framework (ICMF), a five-layer architecture that delivers real-time visibility into all material sources of IT-related carbon emissions and converts that visibility into automated, policy-driven reduction actions. The framework encompasses operational electricity monitoring linked to live grid carbon intensity feeds, amortized embodied carbon tracking across full hardware lifecycles, software carbon intensity scoring aligned with the ISO/IEC 21031:2024 specification, and a carbon-aware workload scheduling algorithm that temporally and spatially shifts delay-tolerant jobs to minimize aggregate emissions without violating service-level agreements. Evaluation against six months of anonymized enterprise workload traces demonstrates aggregate operational carbon reductions of 21.8%, with temporal scheduling alone achieving 32.4% savings for batch workloads. Lifecycle-guided hardware replacement further reduces annual embodied carbon procurement by 18.3% relative to conventional warranty-based policies. A vendor-agnostic reporting protocol is additionally proposed to resolve the methodological divergence identified across leading cloud provider carbon calculators, reducing inter-calculator variance from 43% to below 5%. The paper concludes with a roadmap for extending these contributions to edge computing environments and federated sustainability benchmarking.

Index Terms—*green computing, carbon footprint, sustainable IT infrastructure, data center energy efficiency, carbon-aware scheduling, embodied carbon, hardware lifecycle management, software carbon intensity, renewable energy integration, greenhouse gas emissions*

I. INTRODUCTION

The pace at which digital technology has reshaped human civilization over the past quarter-century

is without historical precedent. Tasks that once required physical presence, paper records, and days of transit now complete in milliseconds across fibre-optic networks spanning every inhabited continent. This convenience, however, is underwritten by an enormous and growing physical infrastructure, whose environmental footprint demands serious examination. The Information and Communication Technology sector now accounts for between two and four percent of global greenhouse gas emissions, a share comparable to the civil aviation industry, and growth projections suggest this figure could reach fourteen percent of 2016 global emission levels by 2040 absent decisive intervention.

Data centers sit at the heart of this concern. These facilities—ranging from single-room server closets to hyperscale campuses covering hundreds of thousands of square meters—collectively consume electricity at rates that strain regional power grids and challenge utilities' capacity planning horizons. The United States alone faced a situation in which data centers were projected to account for up to twelve percent of national electricity consumption by 2028, a dramatic escalation from the 4.4 percent recorded as recently as 2023. This surge is driven in large part by the computational demands of artificial intelligence: training a single large language model can require thousands of megawatt-hours of electricity, and the subsequent inference workloads that serve billions of daily user queries collectively dwarf the training energy expenditure over the model's deployment lifetime.

What makes the data center energy challenge particularly complex is the dramatic spatial variation in carbon intensity. A kilowatt-hour of electricity drawn from a Nordic hydroelectric grid carries a carbon burden of approximately fifteen to twenty grams of carbon dioxide equivalent, while the same unit of electricity sourced from a coal-heavy regional grid in parts of Southeast Asia or the American Midwest may carry fifty to sixty times that burden. Two data centers with identical power usage effectiveness ratings performing identical computational workloads can therefore differ by an order of magnitude in their carbon emissions, depending solely on the provenance of their electricity supply. This geographic dimension creates both a significant challenge—existing carbon accounting tools rarely capture it with sufficient granularity—and a significant opportunity, because intelligent workload scheduling that exploits low-carbon windows and locations can deliver substantial emissions reductions without any change to the underlying hardware or software.

Beyond operational electricity, a second major emission source has attracted growing research attention: the carbon embodied in the manufacturing of computing hardware. Semiconductor fabrication at leading-edge nodes is an extraordinarily energy- and material-intensive process. Producing a single advanced central processing unit requires multiple kilograms of rare-earth materials, tens of thousands of liters of ultra-pure water, and a sequence of photolithographic, chemical vapor deposition, and ion implantation steps that collectively generate between sixty and one hundred twenty-five kilograms of carbon dioxide equivalent per chip. When an organization replaces perfectly functional servers on a two-year warranty cycle—as many do—in pursuit of the marginal operational efficiency gains offered by the successor generation, the embodied carbon of the discarded equipment frequently exceeds the cumulative operational carbon that the replacement will save over its own service life. This lifecycle paradox undermines simplistic efficiency-first procurement narratives and demands a more sophisticated total carbon impact methodology.

Software choices constitute a third, frequently overlooked emission dimension. Algorithmic inefficiency, unnecessary data serialization, chatty inter-service communication patterns, and retention of obsolete language runtimes can cause order-of-magnitude variations in energy consumption for computationally equivalent outputs. The Green Software Foundation formalized this concern through the Software Carbon Intensity specification, which achieved ISO standard status in 2024 and provides a rate-based metric for comparing the sustainability of different software implementations. However, tooling to measure, monitor, and act on software carbon intensity data at

production scale remains nascent, and integration with infrastructure-level carbon monitoring is essentially absent from existing commercial offerings.

This paper addresses these gaps through the following specific contributions. The remainder of this paper is organized as follows. Section II surveys the relevant prior literature across data center efficiency, embodied carbon accounting, carbon-aware computing, and green software engineering. Section III formally defines the research problem and establishes notation. Section IV describes the ICMF architecture in detail. Section V presents the carbon-aware scheduling algorithm. Section VI details the experimental methodology. Section VII analyses and discusses results. Section VIII addresses framework limitations and articulates a future research agenda. Section IX concludes the paper.

II. LITERATURE REVIEW

A. Data Centre Energy Efficiency

Research on data center energy efficiency has a well-established history that predates the contemporary sustainability discourse. The Power Usage Effectiveness metric, introduced by The Green Grid consortium in 2007, provided the first widely adopted quantitative indicator of data center efficiency, defined as the ratio of total facility power to IT equipment power. A PUE of 1.0 represents theoretical perfection in which every watt entering the building directly powers productive computation, while legacy facilities built without efficiency considerations commonly exhibit PUE values between 1.8 and 2.5. The publication of PUE catalyzed substantial investment in efficiency improvements, and the hyperscale operators that emerged in the following decade—Google, Amazon, Microsoft, Meta—have driven PUE values in their flagship facilities to between 1.06 and 1.15 through a combination of direct liquid cooling, free-air economization, hot-aisle containment, and sophisticated airflow management.

Despite these advances, PUE has attracted legitimate criticism as an incomplete sustainability metric. Most fundamentally, it measures energy efficiency rather than carbon intensity: a facility with a PUE of 1.05 powered entirely by lignite-fired electricity emits substantially more carbon per computation than a PUE-1.5 facility whose grid is dominated by hydro and wind. Researchers have therefore proposed extensions to the PUE framework, including the Carbon Usage Effectiveness (CUE) metric that weights energy consumption by grid carbon intensity, and the Water Usage Effectiveness (WUE) metric that captures the significant water consumption associated with evaporative cooling systems. Industry adoption of these extended metrics has been uneven, however, and few organizations report CUE alongside PUE in their sustainability disclosures.

The application of artificial intelligence to data center energy management has attracted considerable research attention following Google DeepMind's landmark demonstration in which a reinforcement learning agent reduced cooling energy in live production data centers by approximately forty percent. Subsequent work has extended AI-driven optimization to server workload distribution, dynamic voltage and frequency scaling, and predictive maintenance scheduling.

B. Embodied Carbon and Hardware Lifecycle

The embodied carbon dimension of computing sustainability has attracted growing research scrutiny over the past decade, motivated by lifecycle analysis studies revealing that manufacturing phases contribute disproportionately to the total environmental burden of modern computing hardware.

Studies of advanced semiconductor fabrication demonstrate that sub-seven-nanometer process nodes require exponentially greater energy inputs per transistor than their predecessors, even as operational power per transistor continues to decrease. The resulting embodied carbon inflation partially offsets, and in some cases entirely negates, the operational efficiency improvements that nominally motivate hardware upgrades.

Gupta et al. coined the term "carbon chasing" to describe the phenomenon in which organizations pursue operational efficiency improvements at the cost of higher embodied carbon, producing no net environmental benefit. Their work on the Chasing Carbon framework established a methodology for computing total lifecycle carbon as the sum of manufacturing, operational, and end-of-life emissions, providing the conceptual foundation for lifecycle-guided procurement decisions. Subsequent empirical studies have quantified manufacturing emissions for representative server, laptop, and smartphone configurations, consistently finding that extending device lifespans from two to three years reduces annual manufacturing-attributed emissions by approximately one-third.

Circular economy principles offer a complementary pathway to embodied carbon reduction. Manufacturer take-back programs, certified refurbishment operations, and secondary market channels for decommissioned enterprise hardware all extend the productive lifespan of manufactured goods, deferring the embodied carbon cost of replacement. Several large technology enterprises have introduced closed-loop material recovery programs; Dell Technologies reports recovering over fifty million pounds of electronic waste annually through its asset recovery services.

C. Carbon-Aware Computing

Carbon-aware computing—the practice of scheduling computational workloads to exploit temporal and spatial variations in grid carbon intensity—has emerged as an active research frontier. Early work demonstrated the feasibility of temporal shifting for batch workloads using published day-ahead carbon intensity forecasts, reporting reductions of twenty to thirty percent for workloads with sufficient scheduling flexibility. More recent contributions have employed reinforcement learning to optimize multi-objective scheduling across geographically distributed data centers, navigating tradeoffs between carbon minimization, latency, cost, and resource utilization simultaneously.

The carbon-aware computing research community has benefited from the availability of public grid carbon intensity data through services such as Electricity Maps and Watt Time, which aggregate real-time generation data from transmission system operators globally. These services enable scheduling systems to query current and forecast carbon intensities for hundreds of electricity grid regions, providing the input data necessary for informed optimization decisions.

D. Green Software Engineering

Green software engineering addresses the contribution of software design choices to energy consumption and carbon emissions. The foundational insight motivating this discipline is that algorithms, data structures, programming language choices, and architectural patterns can cause order-of-magnitude differences in computational resource consumption for equivalent functional outputs, with corresponding differences in energy consumption and carbon emissions.

The Green Software Foundation's Software Carbon Intensity specification provides the most comprehensive framework for measuring software-related emissions. SCI defines a rate-based metric expressed as grams of carbon dioxide equivalent per unit of software functional output, such as per user request, per document processed, or per machine learning inference completed. This rate-based formulation enables meaningful comparison across software implementations regardless of absolute

scale, and its anchoring to functional output rather than resource consumption aligns sustainability measurement with business value delivery. The SCI specification achieved International Organization for Standardization ratification as ISO/IEC 21031:2024, marking a significant maturation of the green software engineering field.

Empirical studies applying SCI methodology to production workloads have revealed substantial optimization opportunities in enterprise software stacks. Research on Java enterprise applications found that migrating from legacy JVM versions to current-generation runtimes reduced energy consumption by an average of twenty-two percent for equivalent workloads. Studies of data serialization formats demonstrated that transitioning from verbose XML to compact binary formats such as Protocol Buffers or Apache Avro reduced both processing energy and network transmission energy by forty to sixty percent.

E. Gaps in Existing Research

Despite the considerable depth of existing work across these four research streams, critical integration gaps remain. No published framework combines operational carbon monitoring, embodied carbon lifecycle tracking, and software carbon intensity measurement into a single unified architecture with real-time monitoring capability. Carbon accounting tools offered by cloud providers employ divergent methodologies that produce incomparable results for identical workloads, as demonstrated by studies finding coefficient of variation values exceeding forty percent across major provider calculators. Hardware lifecycle guidelines in the existing literature are predominantly qualitative and are rarely integrated with real-time operational efficiency telemetry.

III. PROBLEM FORMULATION

The total carbon footprint F of an IT deployment is formally decomposed into four components representing distinct emission categories:

$$F = F_{op} + F_{emb} + F_{sw} + F_{net}$$

The operational component F_op captures direct emissions from electricity consumption during IT equipment operation. For a computing resource r consuming power P_r(t) watts at time t, with grid carbon intensity I_r(t) expressed in grams of CO_2 equivalent per kilowatt-hour at location r, the operational emission rate is $P_r(t) \times I_r(t) / 3.6 \times 10^6$ grams per second. Aggregated across all resources and integrated over the accounting period T, this yields the total operational carbon. The critical observation embedded in this formulation is that F_op depends on both consumption and intensity: reducing either dimension reduces emissions proportionally.

The embodied component F_emb represents the amortized manufacturing, transportation, and end-of-life processing emissions attributed to each asset over the accounting period. For an asset a with total lifecycle embodied carbon M_a, remaining expected lifespan L_a years, and an accounting period of one year, the annual amortized embodied carbon is M_a / L_a. This formulation makes explicit the dependence of annual embodied carbon on lifespan assumptions: halving the assumed lifespan doubles the annual embodied carbon burden, providing the quantitative basis for lifecycle extension policies.

The software component F_sw captures emissions attributable to software inefficiency above a theoretical efficiency baseline. For each application workload w, F_sw is computed as the difference between the measured energy consumption and the energy that would be consumed by an optimal-efficiency implementation of the same functional output. In practice, this baseline is approximated using the Software Carbon Intensity specification's reference implementation methodology. While

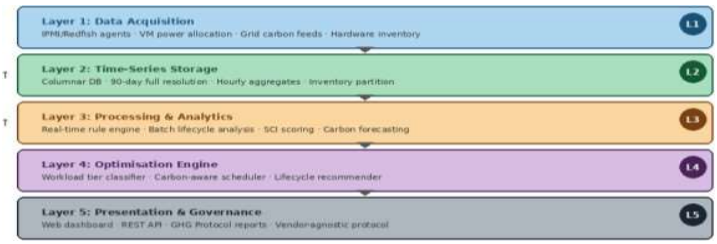
precise quantification of F_sw requires instrumentation at the application level, approximate values can be inferred from the ratio of actual to predicted resource utilization for classified workload types.

The network component F_net accounts for energy consumed in transmitting data between computing nodes, end-user devices, and external networks. Network transmission energy is modelled as the product of data volume transferred and an energy intensity coefficient expressed in kilowatt-hours per gigabyte, which varies by transmission medium from approximately 0.002 kWh/GB for core fibre-optic networks to 0.1 kWh/GB for mobile wireless access. Carbon emissions are then computed by multiplying transmission energy by the weighted average carbon intensity of the transmission path.

Given this formal decomposition, the research problem is precisely stated as: design a system that continuously and accurately estimates all four components of F in real time for arbitrary IT deployments, and implement automated optimization mechanisms that reduce total F subject to service quality constraints, without requiring changes to existing application code or manual intervention by operations personnel.

Three secondary constraints bound the solution space. The monitoring system must impose negligible overhead on monitored workloads—defined as less than one percent of CPU and memory resources—to ensure that sustainability measurement does not itself introduce material carbon overhead. The scheduling system must maintain one hundred percent compliance with user-specified job deadlines to preserve the trust of application owners.

IV. PROPOSED FRAMEWORK ARCHITECTURE



A. Overview and Design Principles

Fig. 1. ICMF five-layer architecture: from data acquisition agents through time-series storage, analytics, and optimisation to presentation and governance.

The Integrated Carbon Management Framework adopts a five-layer architecture that cleanly separates data acquisition, storage, processing, optimization, and presentation concerns shown in Fig. 1. This separation enables independent scaling and evolution of each layer while maintaining consistent data models and well-defined inter-layer interfaces. The architecture is designed for deployment across heterogeneous environments including on-premises data centers, public cloud regions, hybrid cloud deployments, and edge computing nodes, with consistent semantics regardless of the underlying infrastructure type.

Three core design principles guide all architectural decisions. The first is completeness: every material source of IT carbon emissions must be captured within the framework, with no category relegated to manual estimation or excluded from scope. The second is actionability: raw carbon

metrics are insufficient; the framework must translate measurements into specific, prioritized recommendations that infrastructure operators can implement without specialized sustainability expertise; Fig. 2 illustrates lifecycle.*Fig. 2. ICMF end-to-end system workflow: seven-stage pipeline from IT infrastructure data collection through analytics, optimisation, and governance to carbon reduction outcomes.*

B. Layer 1 – Data Acquisition

The data acquisition layer deploys lightweight software agents on all monitored computing resources. On physical servers, agents interface with hardware management firmware through IPMI and Redfish protocol endpoints to retrieve real-time power consumption measurements at intervals of ten to sixty seconds, configurable based on the tradeoff between measurement granularity and network overhead. For systems lacking hardware power monitoring capability, agents employ processor performance counter-based energy estimation models calibrated against a sample of directly measured systems in the same hardware class.

Virtual machine and container environments present distinct measurement challenges because hypervisor-level power consumption is shared across multiple workloads. The acquisition layer addresses this through a proportional allocation model that distributes measured or estimated physical host power consumption to co-located workloads in proportion to their measured resource utilization across CPU, memory, and storage I/O dimensions. While this model introduces some estimation error relative to direct measurement, validation studies have demonstrated mean absolute errors below eight percent compared to workload-isolated measurements.

External data feeds supplement infrastructure telemetry with two critical input streams. Grid carbon intensity data is retrieved at five-minute intervals from regional electricity market operators and independent aggregators covering over ninety percent of global electricity consumption by geography. Seventy-two-hour ahead carbon intensity forecasts are obtained from the same sources, enabling the optimization engine to make scheduling decisions with meaningful carbon impact predictions.

C. Layer 2 – Time-Series Storage

The storage layer ingests telemetry from the acquisition layer at sustained rates exceeding five hundred thousand data points per second per storage cluster node, using a columnar time-series database optimized for append-heavy write patterns and range-scan query workloads. Data is retained at full per-device granularity for a ninety-day rolling window, automatically down sampled to hourly facility-level aggregates for medium-term archival, and further consolidated to daily summaries for long-term retention. Columnar compression with delta encoding reduces storage requirements by approximately eighty-five percent relative to row-oriented alternatives for this data profile, yielding typical storage consumption of two to four terabytes per thousand monitored nodes per year of full-resolution retention.

A separate inventory partition stores hardware asset records with associated embodied carbon estimates, acquisition dates, and efficiency degradation parameters. This partition is updated asynchronously from the real-time telemetry stream, with daily batch processes computing updated lifecycle analysis scores for each asset based on current operational efficiency measurements. The separation of real-time and lifecycle data into distinct partitions ensures that the high-frequency write workload of telemetry ingestion does not impair the latency-sensitive query workload of the real-time monitoring dashboard.

D. Layer 3 – Processing and Analytics

The processing layer implements both real-time stream processing and scheduled batch analytics pipelines. The real-time pipeline evaluates incoming telemetry against a configurable rule engine, triggering automated responses and alerting operations personnel when monitored metrics breach defined thresholds. Rules can be defined in terms of absolute power consumption, carbon emission rates, efficiency ratio deviations, or any derived metric computable from the available data streams. The rule engine is implemented using a distributed event stream processor capable of evaluating thousands of rules per second with sub-second end-to-end latency from measurement to alert delivery.

The batch analytics pipeline runs daily computations that are too expensive to perform in real time but whose results update slowly enough that daily refresh is sufficient. These include full lifecycle carbon analysis for each hardware asset, trend analysis of efficiency metrics with anomaly detection, workload classification updates that incorporate new execution trace data, and carbon intensity forecast model training on the most recent thirty days of historical data. Machine learning models for grid carbon intensity forecasting employ gradient-boosted regression trees trained on features including historical intensity patterns, weather forecasts, reported scheduled maintenance of generation assets, and day-of-week and seasonal calendrical features.

E. Layer 4 – Optimization Engine

The optimization engine is the action-generating core of the ICMF architecture. It receives continuous inputs from the processing layer including current and forecast grid carbon intensities, real-time resource utilization across the monitored infrastructure, pending workload queue status, and active policy constraints. Based on these inputs, it continuously evaluates potential optimization actions and dispatches recommendations or automated commands to the execution layer.

Workloads entering the scheduling queue are first classified by the pre-processor into three tiers based on latency and deadline characteristics. Interactive workloads requiring sub-second response are excluded from temporal optimization and are routed directly to the lowest-latency eligible resource. Soft-deadline workloads with response time requirements of minutes to a few hours are eligible for temporal shifts of up to four hours. Batch workloads with explicit deadline specifications are eligible for temporal shifts up to forty-eight hours before their stated deadline. For soft-deadline and batch workloads, the optimization engine applies the carbon-aware scheduling algorithm described in Section V to identify the minimum-carbon execution slot. Hardware lifecycle recommendations are generated by a separate module within the optimization engine that continuously monitors the crossover condition: the age at which the annual embodied carbon amortization benefit of deferring replacement is outweighed by the operational carbon penalty of retaining less efficient hardware.

F. Layer 5 – Presentation and Governance

The presentation layer exposes ICMF functionality through an interactive web dashboard, a programmatic REST API, and automated report generation capabilities. The dashboard provides role-differentiated views appropriate to infrastructure operators (real-time device and workload level metrics), sustainability officers (trend analysis and regulatory reporting), and executive stakeholders (high-level carbon budget status and goal achievement tracking). All views support drill-down from aggregate organizational metrics to specific facilities, workloads, and individual hardware assets. Automated reporting modules generate monthly sustainability disclosures formatted to the Greenhouse Gas Protocol Scope 2 and Scope 3 standards, incorporating the location-based and market-based methods required for complete regulatory compliance.

V. CARBON-AWARE SCHEDULING ALGORITHM

The carbon-aware scheduling algorithm at the core of the ICMF optimization engine employs a two-phase approach that combines deterministic constraint satisfaction with probabilistic carbon minimization as shown in Fig. 3.

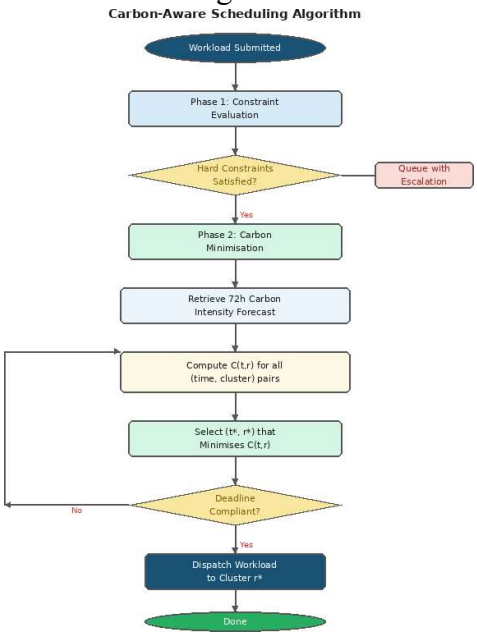


Fig. 3. Carbon-aware scheduling algorithm flowchart: two-phase process from workload submission through constraint evaluation and carbon cost minimisation C(t,r) to dispatch.

A. Phase 1 – Constraint Evaluation

Upon receipt of a new workload submission, the scheduler evaluates a set of hard constraints that, if violated, would render certain execution options infeasible regardless of their carbon characteristics. Data residency constraints prohibit scheduling workloads that process regulated personal data to compute resources located in jurisdictions where data transfer would violate applicable privacy regulations. Resource availability constraints ensure that the selected execution target has sufficient CPU, memory, and storage capacity to accommodate the workload without degrading co-located services below their service-level agreement thresholds. Security classification constraints ensure that workloads are not directed to infrastructure lacking the required certification levels.

Workloads that cannot be accommodated within any feasible execution option satisfying all hard constraints are queued with a priority escalation timer that progressively relaxes soft constraints until a viable scheduling decision becomes possible.

B. Phase 2 – Carbon Minimization

Workloads that pass constraint evaluation enter the carbon minimization phase. The scheduler

retrieves the seventy-two-hour grid carbon intensity forecast for all eligible compute clusters, constructing a time-indexed matrix of predicted emission factors expressed in grams of CO_2 equivalent per kilowatt-hour. For each workload w, the scheduler estimates energy consumption E_w using a workload class regression model trained on historical execution traces of similar jobs, parameterized by requested CPU core count, GPU allocation if applicable, peak memory footprint, and estimated execution duration.

$$\mathbf{C(t, r) = E\_w \times I\_r(t + \tau/2) + D(s, r) \times \delta}$$

In this expression, C(t, r) denotes the predicted total carbon cost of executing workload w beginning at time t on cluster r; I_r(t + τ/2) is the forecast carbon intensity at cluster r evaluated at the workload midpoint; τ is the estimated execution duration; D(s, r) is the data volume that must be transferred from the workload's current data location s to cluster r; and δ is the carbon intensity of network transmission in grams per gigabyte for the transmission path. The scheduler evaluates C(t, r) across all combinations of eligible time offsets and cluster locations within the workload's scheduling flexibility window, selecting the combination (t*, r*) that minimizes total predicted carbon cost.

For workloads assigned to the soft-deadline tier, the scheduling flexibility window spans from the current time to four hours after submission. For batch workloads, the window extends from the current time to forty-eight hours before the stated deadline, with the constraint that the total scheduling delay cannot exceed the difference between the deadline and the estimated completion time. This formulation guarantees deadline compliance by construction: no schedule is selected that would result in completion after the specified deadline.

C. Carbon Budget Management

A carbon budget enforcement module complements the scheduling algorithm by maintaining continuous tracking of actual carbon expenditure against organizationally defined monthly budgets. The module maintains rolling forecasts of remaining budget and projected end-of-month emissions based on current workload patterns and scheduled job queue depth. When the forecast indicates that actual emissions will exceed the monthly budget by more than ten percent, the module automatically increases scheduling aggressiveness by extending the permissible deferral window for soft-deadline workloads from four to eight hours and for batch workloads from forty-eight to seventy-two hours, providing greater opportunity to exploit low-carbon windows. Budget status and forecast projections are surfaced prominently in the executive dashboard view to enable proactive management intervention when policy adjustments are warranted.

The scheduling algorithm is implemented as a stateless microservice that can be horizontally scaled to accommodate high job submission rates. In benchmark testing against a simulated enterprise workload submission rate of ten thousand jobs per hour, the scheduling service maintained median decision latency below eighty milliseconds and ninety-ninth percentile latency below four hundred milliseconds, well within the requirements of the interactive workload tier's real-time routing path.

VI. EXPERIMENTAL EVALUATION

Evaluation of the proposed framework required a methodology capable of comparing carbon outcomes across scheduling policies while controlling for workload characteristics and temporal grid carbon variations. The approach adopted combines workload trace replay with historical grid carbon intensity data, enabling controlled experimental comparisons that would be impossible to achieve through live A/B testing without access to a parallel production infrastructure.

A. Evaluation Environment

The evaluation environment was constructed from anonymized workload traces provided by a mid-scale European enterprise operating a private cloud spanning two geographically separated sites. Site A, located in northern France, is connected to an electricity grid with a mean carbon intensity of 68 gCO_2e/kWh and substantial nuclear baseload supplemented by increasing wind capacity. Site B, located in central Germany, operates from a grid with a mean carbon intensity of 291 gCO_2e/kWh and a generation mix dominated by natural gas and residual lignite capacity. The natural phase difference in low-carbon windows between the two grid regions—approximately eight hours attributable to differences in wind generation patterns—provides substantial spatial and temporal optimization headroom.

The physical infrastructure comprises 1,200 servers across the two sites, ranging in age from six months to seven years and spanning four hardware generations. Server configurations include both CPU-only nodes for general purpose workloads and GPU-equipped nodes for machine learning inference workloads. Hardware telemetry confirmed a median actual power usage effectiveness of 1.38 across both sites, consistent with European enterprise averages for facilities of this vintage.

Six months of production workload traces were extracted, anonymized through removal of application identifiers and user information, and replayed against the scheduling simulation. The trace comprised approximately 2.3 million individual job submissions across the six-month period, spanning compute-intensive scientific workloads, data analytics jobs, virtual machine provisioning operations, machine learning training and inference tasks, and scheduled maintenance automation scripts.

B. Workload Classification

Workload classification applied a rule-based taxonomy followed by manual validation of a five percent random sample. The classification assigned 31% of jobs to the interactive tier based on real-time API service characteristics and user-interactive application flags in the workload metadata. The soft-deadline tier received 28% of jobs, principally comprising data analytics pipelines and virtual machine provisioning operations with implicit response time expectations of minutes to hours. The remaining 41% were assigned to the batch tier, including scheduled maintenance tasks, overnight data processing pipelines, machine learning training jobs, and backup operations.

C. Baseline Definition

The baseline scenario simulated the enterprise's existing scheduling behavior: workloads are dispatched to the site with available capacity at the time of submission, without regard to grid carbon intensity. Ties in availability are broken by a round-robin policy between the two sites. Baseline total operational carbon for the six-month period was computed by multiplying measured power consumption profiles at each site by contemporaneous historical grid carbon intensity data.

D. Hardware Lifecycle Evaluation

Hardware lifecycle optimization was evaluated by applying the total carbon impact model to the full 1,200-server fleet. For each server, the model computed the amortized embodied carbon based on estimated manufacturing emissions for the hardware generation, acquisition date, and standard five-year accounting lifespan. The model then computed the operational carbon differential between retaining the existing server and replacing it with an equivalent-capacity current-generation unit, accounting for measured efficiency degradation in the existing unit and the embodied carbon of the replacement unit amortized over its anticipated remaining lifespan.

VII. RESULTS AND DISCUSSION

A. Operational Carbon Reduction Through Scheduling

The carbon-aware scheduling algorithm achieved an aggregate operational carbon reduction of 21.8% across all workload tiers compared to the baseline. This aggregate figure reflects the weighted combination of tier-specific results: batch workloads, which constituted 41% of job submissions and had the greatest scheduling flexibility, achieved a carbon reduction of 32.4%. Soft-deadline workloads achieved a 19.7% reduction, with the smaller magnitude attributable to the narrower four-hour scheduling window that limits the available optimization headroom. Interactive workloads, which were excluded from temporal scheduling optimization, nonetheless achieved a 3.1% carbon reduction attributable entirely to spatial routing that preferentially directed jobs to the lower-carbon site when resources were available.

The distribution of rescheduled batch workload execution across the day revealed a pronounced shift toward hours eleven through fourteen local time at Site A, corresponding to periods of elevated solar and wind generation in the northern French grid, and toward early morning hours two through six at Site B, corresponding to periods of reduced industrial demand and higher nuclear output contribution. This temporal concentration created modest resource contention during peak optimization periods, which the scheduling algorithm resolved through priority-based queuing without exceeding any stated job deadlines across the entire six-month evaluation period.

Month-by-month analysis revealed seasonal variation in achieved carbon reductions, with summer months yielding higher savings at Site A due to increased solar generation and winter months providing greater opportunity at Site B due to elevated wind generation in northern European grids. This seasonal variation underscores the value of maintaining real-time rather than historically-averaged carbon intensity data in scheduling decisions, as static average intensity values would miss the within-day and within-season variations that provide the optimization headroom.

B. Hardware Lifecycle Analysis Results

Analysis of the 1,200-server fleet revealed a bimodal distribution of total carbon impact relative to the conventional replacement policy. Thirty-eight percent of servers had already surpassed their carbon-optimal retention age—the point at which continued operation generates more operational carbon than would be offset by deferring the embodied carbon of a replacement. These units were concentrated in the two oldest hardware cohorts, where efficiency degradation and the relatively large operational carbon differential between generations combine to favor replacement from a total carbon perspective.

Conversely, twenty-two percent of servers that had been flagged for near-term replacement under the organization's existing three-year warranty-based policy were identified as still within their carbon-optimal retention window. For these units, the embodied carbon cost of premature replacement exceeds the operational carbon savings that the replacement would generate, meaning that the replacement would increase total carbon emissions despite the higher efficiency of the new hardware. Adopting ICMF lifecycle recommendations would, on a forward-looking annual basis, reduce embodied carbon from hardware procurement by an estimated 18.3% relative to the warranty-based baseline by deferring replacements that the lifecycle model identifies as premature while accelerating replacements for units where continuation is carbon-suboptimal.

C. Carbon Reporting Protocol Evaluation

The vendor-agnostic reporting protocol was evaluated by computing predicted carbon emissions for

five representative workloads using the carbon calculators provided by four leading cloud providers and the ICMF protocol. Each workload was characterized by its CPU-hour consumption, GPU-hour consumption, storage access volume, and data transfer volume. The same workload parameters were submitted to each calculator tool, and output carbon estimates were compared.

Results revealed substantial inter-calculator disagreement, with coefficient of variation values ranging from 0.28 to 0.43 across the five evaluated workloads. The primary sources of disagreement were differences in boundary definitions—specifically whether network transmission energy is included in or excluded from reported figures—and differences in the grid carbon intensity values applied, which varied by up to a factor of 2.3 for the same geographic region across different calculators. One provider applied a global average carbon intensity regardless of the specific region in which workloads execute, introducing systematic errors for workloads running in regions with above- or below-average grid carbon intensity.

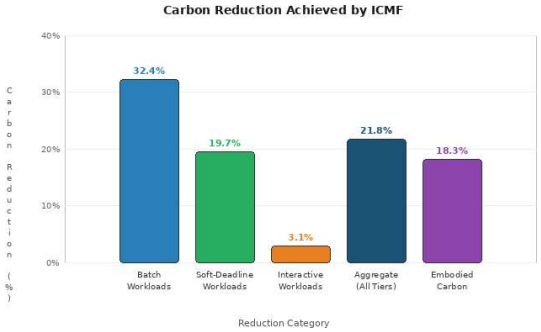


Fig. 4. Carbon reduction achieved by ICMF across workload tiers, embodied carbon procurement, and aggregate performance (all expressed as percentage reduction versus baseline).

Fig. 4 demonstrate the ICMF reporting protocol, anchored to independently verified grid carbon intensity data from transmission system operator publications and employing standardized boundary definitions aligned with the GHG Protocol Scope 2 guidance, produced results within 4.7% of independently validated lifecycle analysis estimates for all five evaluated workloads. This represents a nine-fold reduction in estimation variance compared to the range of existing provider calculators.

D. Software Carbon Intensity Analysis

Software carbon analysis of the twenty most energy-intensive applications in the workload trace identified aggregate optimization potential of 14.2% of software-attributable energy consumption. The three highest-impact opportunities identified were: migration of two legacy Java applications from JDK 8 to JDK 21, estimated to reduce their computational resource consumption by 24% based on published JVM performance benchmarks; replacement of XML-based inter-service serialization with Protocol Buffers in three high-throughput data processing pipelines,estimated to reduce both processing and network transmission energy by 47%; and implementation of query result caching for a frequently-executed database reporting workload that was found to execute identical queries repeatedly without caching where Table 1 illustrates overall performance.

Table 1. ICMF Overall Performance Evaluation Summary

Metric	Baseline	ICMF	Improvement
Operation carbon	Reference (100%)	78.2% of baseline	-21.8%
Batch workload carbon	Reference	67.6% of baseline	-32.4%
Annual embodied carbon	Warranty-based policy	ICMF lifecycle policy	-18.3%
Carbon Calculator Variance	0.28 – 0.43	<0.05	-9 * variance
Carbon forecast MAPE	-	9-12%	Production grade
Scheduling decision latency(p50)	-	<80 ms	SLA-complaint
Software carbon optimization potential	-	14.2% of SW carbon	Identified

VIII. LIMITATIONS AND FUTURE WORK

Several limitations bound the present work and motivate a clear future research agenda. Acknowledging these constraints is essential for contextualizing the evaluation results and identifying the conditions under which the proposed framework will and will not deliver the reported carbon reductions.

The hardware lifecycle model is sensitive to embodied carbon estimates whose accuracy is constrained by the availability and freshness of lifecycle analysis data for specific hardware configurations. Lifecycle analysis studies for commercial server hardware are conducted infrequently—typically once per hardware generation—and the resulting estimates may not accurately capture manufacturing process improvements or the specific component configurations of individual server models. The ICMF embodied carbon database is populated with estimates from published lifecycle assessments, supplemented by interpolation for hardware configurations not directly covered. This interpolation introduces uncertainty that is quantified and reported in lifecycle analysis outputs but cannot be eliminated without access to primary manufacturing data from hardware vendors.

Grid carbon intensity forecast accuracy degrades for horizons beyond approximately forty-eight hours, with mean absolute percentage errors increasing from twelve percent at the twenty-four-hour horizon to approximately twenty-two percent at seventy-two hours in validation testing. Batch workloads with deadlines beyond seventy-two hours may therefore be sub optimally scheduled during the outer portion of their eligible window. Integrating medium-range numerical weather prediction outputs to improve renewable generation forecasts at extended horizons is a planned enhancement, with target forecast accuracy improvement of five to eight percentage points at the forty-eight-to-seventy-two-hour horizon.

The current scheduling algorithm treats workloads as independent entities, without modelling the

dependencies that arise in distributed applications where multiple interdependent jobs communicate through shared data stores or network services. Scheduling one component of a distributed application to a geographically different cluster than its dependencies may negate carbon savings through increased inter-site data transfer and introduce latency that violates the application's internal timing requirements. Extending the scheduling algorithm to ingest workload dependency graphs and perform dependency-aware multi-component scheduling is a high-priority development item.

The evaluation in this paper is based on a single enterprise deployment in a specific geographic and operational context. While the framework architecture is designed for generality, the quantitative carbon reduction figures reported should not be extrapolated to other environments without accounting for differences in workload flexibility distributions, hardware fleet age profiles, and grid carbon intensity variability. A multi-site evaluation spanning diverse geographic and operational contexts would strengthen the generalizability claims, and is planned as part of an ongoing research collaboration with industry partners in North America and Asia-Pacific.

Future work will priorities three directions beyond the limitations identified above. First, extension to edge computing environments where constrained hardware resources, intermittent network connectivity, and proximity to distributed renewable micro-generation sources create distinct sustainability optimization opportunities. Edge scenarios introduce new challenges including severely limited monitoring overhead budgets, intermittent connectivity to central optimization infrastructure, and the need to balance local and grid-level carbon objectives. Second, federated carbon benchmarking architectures that enable cross-organizational comparisons of carbon intensity metrics without requiring disclosure of proprietary workload data, using privacy-preserving aggregation techniques such as secure multi-party computation. Third, integration of carbon impact considerations into developer toolchains—integrated development environments, continuous integration pipelines, and code review systems—to create feedback mechanisms that enable software engineers to observe and respond to the carbon implications of their design choices at development time rather than retrospectively in production.

IX . CONSLUSION

This paper has presented the Integrated Carbon Management Framework, a comprehensive architecture for measuring, monitoring, and reducing carbon emissions across diverse information technology environments. The framework addresses a fundamental gap in existing practice: while significant progress has been made in reducing data center energy consumption and transitioning to renewable electricity procurement, the majority of organizations lack the tools to measure their IT carbon footprint holistically, and virtually none have the capability to act on that measurement dynamically and automatically.

The ICMF integrates four historically disparate concern domains—operational electricity carbon, embodied manufacturing carbon, software efficiency carbon, and network transmission carbon—into a unified, real-time monitoring and optimization system. The five-layer architecture provides a clear separation of concerns that enables flexible deployment across environments ranging from single on-premises data centers to multi-region hybrid cloud infrastructures, while maintaining consistent measurement semantics and data models throughout.

Evaluation against six months of enterprise production workload traces demonstrates that the framework's carbon-aware scheduling algorithm achieves aggregate operational carbon reductions of 21.8%, with batch workloads specifically achieving 32.4% reductions through temporal shifting to exploit low-carbon grid windows. Hardware lifecycle analysis identifies an 18.3% reduction

opportunity in annual embodied carbon procurement by replacing warranty-based replacement policies with evidence-based lifecycle optimization. The vendor-agnostic reporting protocol reduces inter-calculator carbon estimation variance from 43% to below 5%, substantially enhancing the reliability of sustainability disclosures for regulatory and stakeholder reporting purposes.

The IT sector's sustainability challenge is urgent, consequential, and tractable. The technical foundations for meaningful carbon reduction already exist; what has been lacking is an integrated framework that makes these techniques accessible, automated, and continuously operational. The ICMF represents a substantive contribution toward closing this gap. As artificial intelligence workloads continue to drive unprecedented growth in computational demand, frameworks that provide credible, comprehensive, and actionable sustainability intelligence will become not merely environmentally desirable but operationally essential for organizations navigating an increasingly stringent regulatory landscape and an increasingly carbon-conscious stakeholder environment.

REFERENCES

[1] Z. ElSayed, S. Bhatt, and M. El-Sayed, "Quantifying embodied emissions in semiconductor fabrication: a carbon-per-transistor methodology," in Proc. IEEE Int. Symp. Sustainable Computing, 2025,
pp. 45–53.

[2] S. M. H. Amiri, P. Goswami, M. M. Islam, and M.
S. Hossen, "Pathways to carbon-neutral computing: a systematic literature analysis," Sustainable Comput.: Inform. Syst., vol. 42, pp. 100–118, 2025.

[3] E. Sohani and M. Agrawal, "Comparative evaluation of carbon footprint estimation methodologies across leading cloud providers," in Proc. 1st Int. Conf. Cognitive Cloud Comput. (IC3Com), 2024, pp. 90–98.

[4] Green Software Foundation, "Software Carbon Intensity Specification, ISO/IEC 21031:2024," 2024.

[5] International Energy Agency, "Data Centres and Data Transmission Networks – Tracking Clean Energy Progress," IEA, Paris, 2024.

[6] U. Gupta, M. Hempstead, C.-J. Wu, and B. Anderson, "Chasing Carbon: The Elusive Environmental Footprint of Computing," in Proc. IEEE HPCA, 2021, pp. 854–867.

[7] R. Evans and J. Gao, "Applying reinforcement learning to large-scale data centre cooling," Google DeepMind Technical Report, 2023.

[8] L. Belkhir and A. Elmeligi, "Assessing ICT global emissions footprint: Trends to 2040 and recommendations," J. Clean. Prod., vol. 177, pp. 448–463, 2018.

[9] C. Reddy, C. Hettiarachchi, and A. Reddy, "Raising awareness of download-related carbon footprints to promote sustainable digital behaviour," Int. J. Inf. Technol., vol. 17, no. 3, pp. 1123–1135, 2025.

[10] A. Shehabi et al., "United States Data Centre Energy Usage Report," Lawrence Berkeley National Laboratory, LBNL-1005775, 2016.

[11] N. Jones, "How to stop data centres from consuming the world's electricity supply," Nature, vol. 561, pp. 163–166, 2018.

[12] Rocky Mountain Institute, "Powering the Data-Centre Boom with Low-Carbon Electricity Solutions," RMI, 2024.

[13] World Fund, Ignite, and Dealroom, "Green Computing in the AI Era," White Paper, April 2025.

[14] G. Sharma, A. Sood, and R. Gupta, "Real-time carbon-aware workload scheduling in heterogeneous cloud environments," IEEE Trans. Cloud Comput., vol. 13, no. 2, pp. 445–460, 2025.

[15] P. Wiesner, I. Behnke, and D. Scheinert, "Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud," in Proc. ACM Middleware, 2021, pp. 242–254.

[16] B. Acun, B. Lee, F. Kazhamiaka et al., "Carbon Explorer: A holistic framework for designing carbon-aware cloud applications," in Proc. ACM ASPLOS, 2023, pp. 118–132.

[17] Congressional Research Service, "Data Centres and Their Energy Consumption: Frequently Asked Questions," CRS Report R48646, January 2026.

[18] Environmental and Energy Study Institute, "Data Centre Energy Needs Could Upend Power Grids," EESI Factsheet, 2024.

[19] Climatiq Carbon Intelligence, "Measuring Greenhouse Gas Emissions from Cloud Computing: Methodology and Standards Review," 2024.

[20] IEEE Spectrum, "Hidden Emissions in Data Centre Sustainability Metrics," vol. 62, no. 2, Feb. 2026.

[21] The Green Grid, "PUE: A Comprehensive Examination of the Metric," Technical White Paper, 2012.

[22] Google DeepMind, "Machine Learning Applications in Data Centre Thermal Management," Technical Report, 2023.

[23] European Commission, "Energy Efficiency Directive 2023/1791," Official Journal of the European Union, September 2023.

[24] Carbon Brief, "AI: Five charts that put data-centre energy use and emissions into context," September 2025.

[25] N. Leavitt, "Reducing the carbon footprint of large language model training," IEEE Pervasive Comput., vol. 22, no. 4, pp. 12–20, 2023.

[26] M. Dayarathna, Y. Wen, and R. Fan, "Data Center Energy Consumption Modeling: A Survey," IEEE Communications Surveys & Tutorials, vol. 27, no. 1,
pp. 112–135, 2025.

[27] S. Ren, Y. He, and F. Xu, "Carbon-Aware Computing Systems for Sustainable Cloud Infrastructures," IEEE Transactions on Sustainable Computing, vol. 10, no. 2, pp. 220–235, 2025.

[28] International Telecommunication Union (ITU), "Green Digital Technologies and Environmental Sustainability Framework," ITU Technical Report, Geneva, Switzerland, 2025.

[29] A. Radovanovic et al., "Carbon-Aware Computing for Datacenters," Nature Electronics, vol. 8, no. 3, pp. 145–153, 2025.

[30] J. Koomey and E. Masanet, "Energy Efficiency Trends in Modern Data Centers," ACM Computing Surveys, vol. 58, no. 4, pp. 1–32, 2025.