GreenSys

UNIVERSITY of York

CEDAR

Carbon Efficient Dynamic Allocation and Routing for Agentic LLM Inference

Amit More
University of York

Tarique Anwar
RMIT University

Poonam Yadav
University of York

April 27–30, 2026 · Edinburgh, Scotland · ACM Workshop on Systems & Methods for Sustainable Large-Scale AI

More, Anwar & Yadav (2026) · GreenSys '26, EuroSys · https://dl.acm.org/doi/10.1145/3802973.3804457

01

The AI Energy Crisis

Rapid AI growth is colliding with global energy constraints

415 TWh

Global data centre electricity consumed in 2024 [1,2]

945 TWh

Projected by 2030 — nearly 3% of global electricity [1]

10×

More electricity per LLM query vs a traditional web search [3]

The Growing Scale

Data centres consumed 415 TWh in 2024, growing 12% per year for five years [1]. Projected to reach 945 TWh by 2030. US AI-specific servers will grow from 53–76 TWh (2024) to 165–326 TWh by 2028 [2].

AI Inference Demand

- AI inference market: $106B (2025) → $255B (2030) [4]
- LLMs power coding assistants, enterprise pipelines, clinical AI, and scientific research
- Demand runs 24/7 across global infrastructure

[1] IEA (2025). Energy Demand from AI. Technical Report. [2] Shehabi et al. (2024). United States Data Center Energy Usage Report. LBNL-2001552. [3] UNRIC (2025). Artificial Intelligence: How Much Energy Does AI Use? [4] MarketsandMarkets (2025). AI Inference Market Size & Growth Analysis.

The Problem: Single-Objective LLM Serving

Production stacks optimise for performance alone carbon and cost are afterthoughts

Single-Objective Focus

TensorRT-LLM, SGLang, vLLM optimise only latency/throughput. No joint cost, carbon, or priority-aware routing in any production system[1,2,3].

Head-of-Line Blocking

Inflates average latency by up to 5.3×. Prefill-decode interference causes generation stalls degrading overall responsiveness [4].

Infrastructure Fragmentation

~28% throughput underutilisation. Systems rarely exploit real-time carbon intensity variation across regions and time [5].

Wrong Carbon Metrics

Naive per-request metrics conflate sunk carbon (baseline infra) with marginal carbon (schedulable emissions) optimizing the wrong target [6].

Central question: Can queue-level, multi-objective control jointly minimise latency, cost, and carbon for agentic LLM workloads?

[1] Kwon et al. (2023). vLLM. SOSP '23 [2] Shen et al. (2024). SGLang. NeurIPS '24 [3] NVIDIA (2024). TensorRT-LLM Library [4] Palke et al. (2024). Queue Mgmt for SLO-Oriented LLM Serving. SoCC '24 [5] Goel et al. (2025). Breaking LLM Serving Silos. arXiv:2503.22562 [6] Chien et al. (2024). The Sunk Carbon Fallacy. SoCC '24

Agentic Workloads Amplify the Challenge

A single coding task generates 10+ sequential LLM calls creating unique scheduling opportunities

Mixed-Criticality Request Tiers

HIGH — Interactive Completion

SLO: 500 ms SLO | 30% of requests

Autocomplete, code generation

MEDIUM — Code Refactoring

SLO: 2 s SLO | 30% of requests

Refactoring, code review

LOW — Analysis / Test Gen

SLO: 5 s SLO | 40% of requests

Batch test gen, doc processing

Key Scheduling Insights

- 40% of agentic requests are LOW priority deferrable for cost & carbon savings [1]
- 37% KV cache recomputation overhead under bursty agentic loads [2]
- Carbon intensity varies 3× across regions and throughout the day [3,4]
- Queue-level control exposes scheduling slack invisible to per-request routing
- Current systems treat all requests independently missing session-level SLO slack
- Opportunity: route deferrable steps to cheaper/greener resources without harming QoS

[1] Gartner (2025). Multi-Agent System Inquiries Surge 1,445%. Industry Report. [2] Goel et al. (2025). Breaking the Silos of LLM Inference Serving. arXiv:2503.22562. [3] WattTime (2025). Real-Time Grid Carbon Intensity Data API. [4] Electricity Maps (2025). Real-Time Electricity Intensity API.

THE RESEARCH QUESTION

LLM calls dominate cost

Agentic pipelines issue 10s–100s of LLM calls per task — routing quality directly impacts the bill

Carbon varies 3× by region

Grid carbon intensity swings 3× across AWS regions and throughout the day — timing matters

No joint optimiser exists

Current routers target latency OR cost — never all three objectives simultaneously

Can a queue-level, multi-objective controller jointly minimise

Latency · Cloud Cost · Carbon

for heterogeneous agentic LLM inference workloads —

without unacceptable degradation of any single objective?

CEDAR answers: Yes.

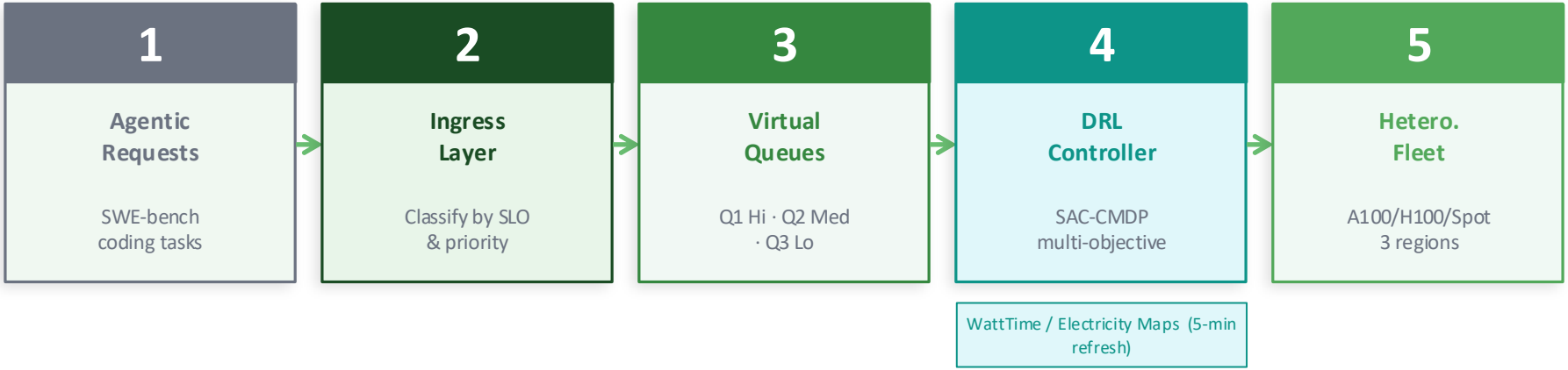
Queue-aware CMDP + SAC controller with live carbon signals

26% cost · 27% carbon · p95 = 0.88 s

04

Introducing CEDAR

A queue-aware, multi-objective control framework for agentic LLM inference



Key Contributions

01 First joint optimiser

Simultaneous minimisation of latency, cloud cost, and marginal carbon at queue level — no prior system does this for agentic LLM inference

02 CMDP + SAC controller

Constrained MDP formulation with SLO slack guards; Soft Actor-Critic learns safe routing under steady, heterogeneous workloads

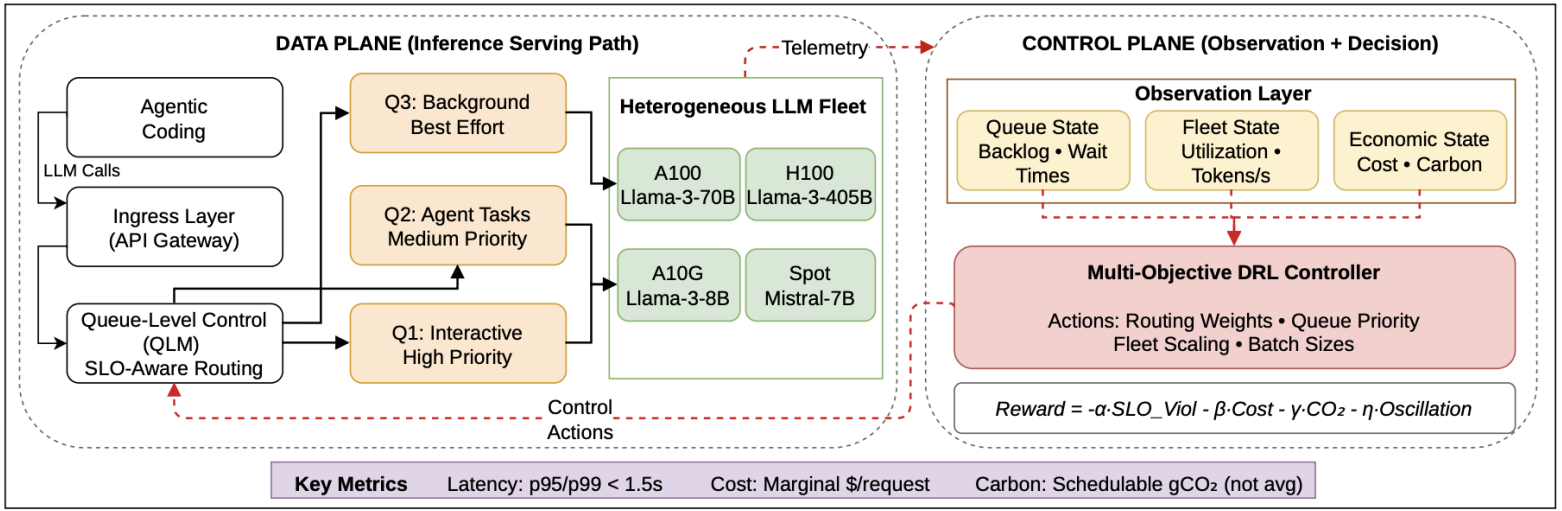
State $s = \{i_t, o_t, \text{latency budget}, \text{queue delay}, CO_2, \text{cost}\}$
Action $a = (\text{region } r, \text{model } m)$

03 Trace-driven results

26% cost reduction, 27% carbon reduction, p95 latency 0.88 s on SWE-bench-derived traces across 3 AWS regions

CEDAR Architecture

CEDAR architecture and control flow. Agentic pipelines generate mixed criticality LLM invocations. The ingress layer classifies requests into virtual queues. The controller observes queue level state, fleet telemetry, and carbon/pricing signals to decide routing and scaling actions over a heterogeneous, geo distributed fleet.



CMDP Controller

Soft Actor-Critic agent jointly optimising latency, cost, and carbon (Conservative Q-Learning)

Reward: $R = -\alpha \cdot SLO\_Viol - \beta \cdot Cost - \gamma \cdot CO_2(marginal) - \eta \cdot Oscillation$

Observations (State)

Queue

Backlog · SLO slack · wait time per tier

Fleet

GPU util · tokens/s per instance

Carbon

Grid intensity c_r(t) · live pricing

SAC Controller

Constrained MDP (CMDP)

SLO slack guards ensure no tier is sacrificed for cost savings

Algorithm: Soft Actor-Critic (SAC)

Actions (every 10 s)

- Routing weights
- Queue priority order
- Fleet scaling
- Batch sizes

QoS Constraints:

HIGH ≤ 500 ms SLO	MED ≤ 2 s SLO	LOW ≤ 5 s SLO

Minimize carbon and cost but never let latency SLOs be violated beyond a threshold.

Experimental Setup

Trace-driven discrete-event simulation with realistic GPU throughput profiles

Fleet Configuration

- 4× A100-80GB us-east-1 — Llama-3-70B
- 4× H100-80GB us-west-2 — Llama-3-405B
- 4× A100-40GB eu-west-1 — Llama-3-8B
- 12 GPU instances · 3 AWS regions
- Routing decisions every 100 ms

Workload

- 10,000 requests per experiment
- 30% HIGH · 30% MED · 40% LOW
- Poisson arrivals at 50 req/s mean
- 3× synthetic bursts (10 s) every 5 min
- Log-normal token lengths (μ=6.5, σ=1.2)

Carbon Signals

- WattTime API at 5-min granularity
- Jan–Mar 2025 real carbon traces
- Covers diurnal & weather-driven variation
- Marginal (not average) carbon attribution
- Electricity Maps for EU region

Baselines (same fleet & carbon traces):

Round Robin	Least Loaded	Performance-Only
Naive equal distribution	Routes to lowest utilisation	Optimises latency; no carbon/cost

Results: Overall Performance

CEDAR vs baselines — 26% cost reduction, 27% carbon reduction, competitive latency

System	p95 Latency	SLO Violation	Cost / Request	gCO_2 / Request	Oscillation
CEDAR	**0.88 s**	**4.3%**	**$0.00189**	**67.8**	**0.19**
Round Robin	1.58 s	22.3%	$0.00245	95.7	0.51
Least Loaded	1.21 s	12.8%	$0.00215	79.3	0.42
Performance-Only	0.76 s	3.2%	$0.00256	93.2	0.29

26%
Cost Reduction
vs Performance-Only

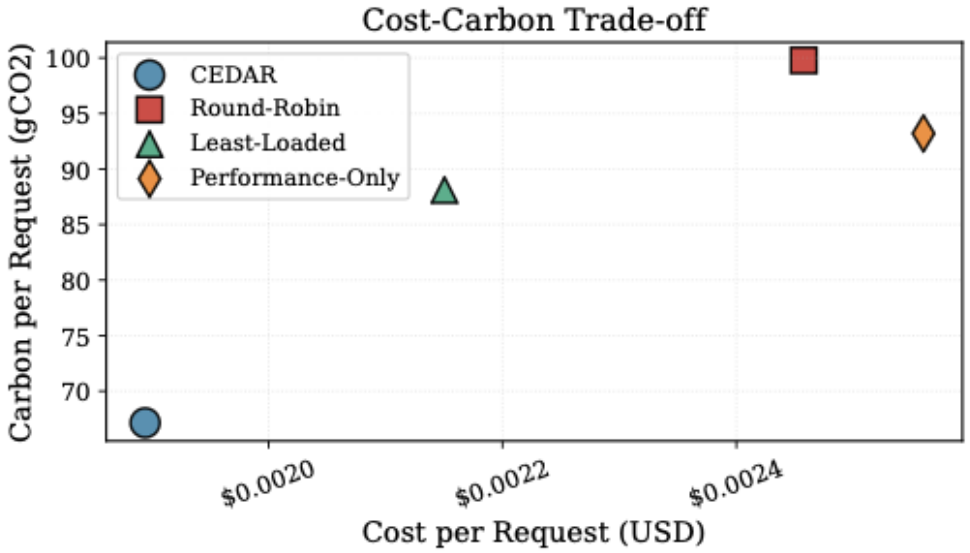
27%
Carbon Reduction
vs Performance-Only

5×
Fewer SLO Violations
vs Round Robin (4.3% vs 22.3%)

0.88s
p95 Tail Latency
+16% vs best — competitive

Carbon Cost Trade Off

CEDAR achieves the best balance of cost and carbon across all strategies



Trade-Off Analysis

CEDAR

Lowest cost and lowest carbon per request

Perf-Only

Higher cost and carbon

Round Robin

Highest carbon emissions

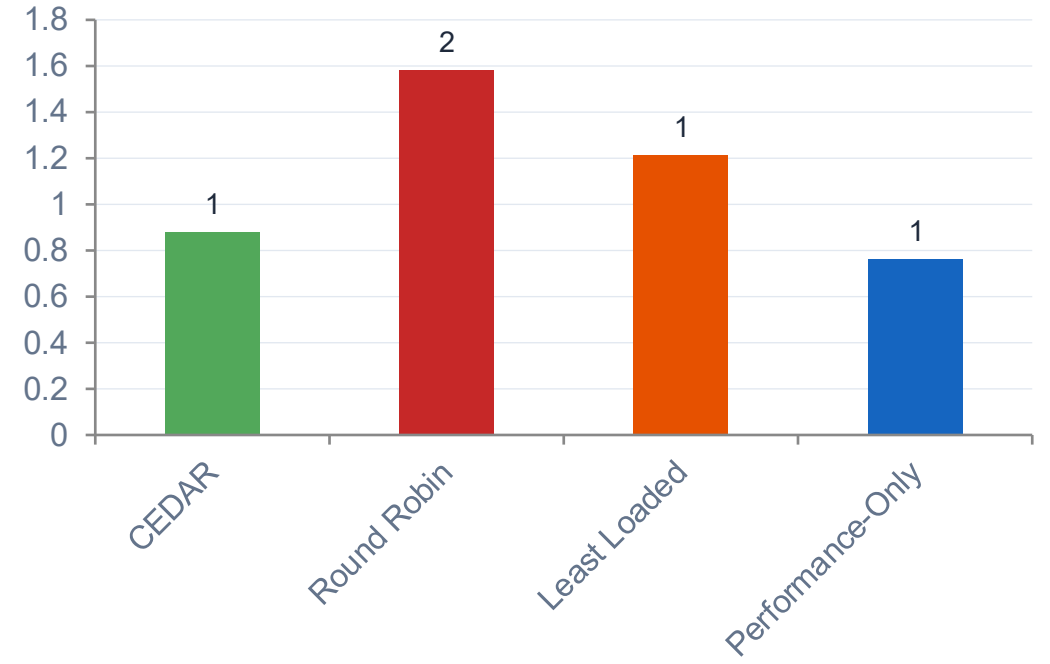
Least Loaded

Better than RR but not optimal

Key insight: CEDAR sits in the optimal region, minimizing both cost and carbon compared to all baselines.

Results: Latency Preservation

Carbon-aware optimisation does not introduce uncontrolled tail amplification



SLO Violation Rate

4.3%	12.8%	22.3%	3.2%
CEDAR	Least Loaded	Round Robin	Perf-Only

Latency Analysis

Only +16% overhead

CEDAR (0.88 s) vs Perf-Only (0.76 s) full carbon awareness at minimal latency cost

44% better p95

vs Round Robin (1.58 s); 27% better than Least Loaded (1.21 s)

5× fewer SLO violations

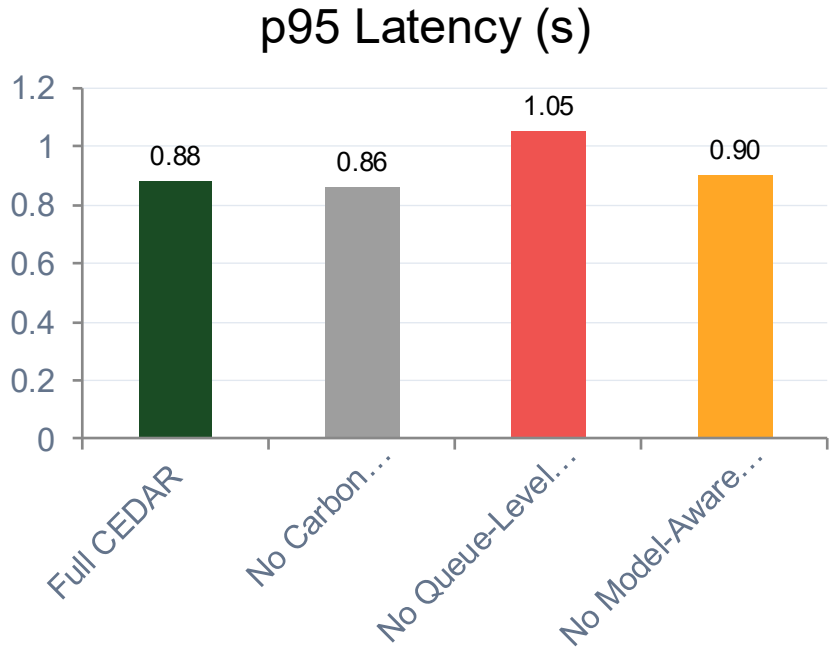
4.3% (CEDAR) vs 22.3% (Round Robin) tail latency well controlled

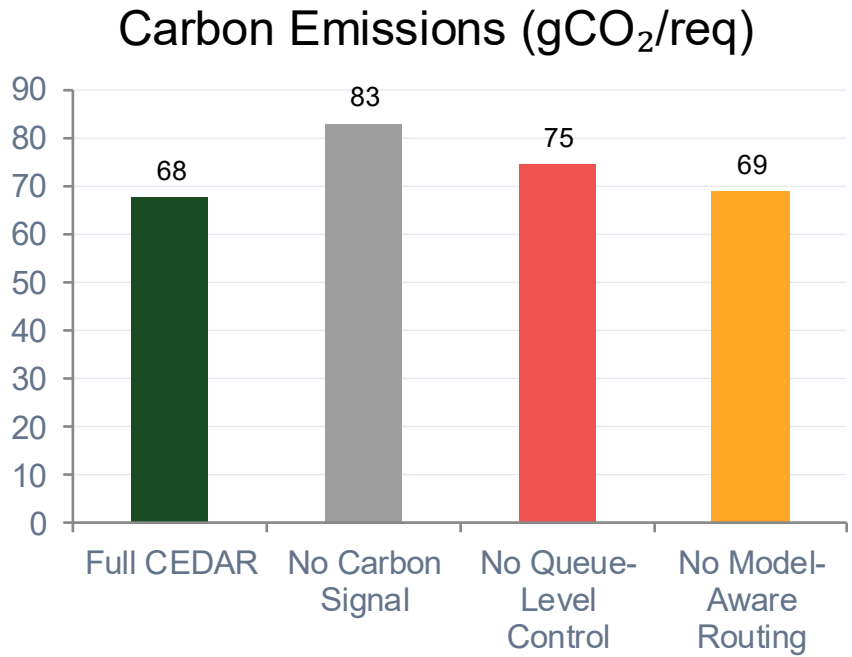
63% less oscillation

Routing stability: 0.19 (CEDAR) vs 0.51 (Round Robin)

Component Analysis: What Drives the Gains?

Each mechanism contributes independently removing any degrades at least one objective





No Carbon Signal → +22% emissions · No Queue Control → +19% latency & +14% cost · No Model-Aware Routing → +9% cost

Limitations & Future Work

Current Limitations

- Simulation only live vLLM cluster deployment pending
- Synthetic workload Azure/production traces not yet used
- SAC training stabilisation under non-stationary demand
- Short evaluation horizon (hours, not 24h–7d)

Future Directions

- Live vLLM deployment measure scheduling overhead in production
- Real LLM traces (Azure conversation datasets & serving logs)
- SAC training stabilisation under non-stationary demand
- Long-horizon evaluation: 24h/7-day workloads
- Spot capacity and multi-tenant fleet scaling
- Integrate live DCGM telemetry directly into the CEDAR observation layer

Acknowledgements

Supported by EPSRC & DSIT: EP/X040518/1, EP/Y037421/1, EP/Y019229/1

DOI: 10.1145/3802973.3804457 · GreenSys '26, April 27–30, Edinburgh

Conclusion

CEDAR demonstrates that queue-level, carbon-aware routing can jointly optimise latency, cost, and marginal emissions for agentic LLM inference without unacceptable QoS degradation.

26%

Cost Reduction

vs Performance-Only

27%

Carbon Reduction

vs Performance-Only

0.88s

p95 Tail Latency

competitive with best

Key Insight: The queue not the individual request is the natural control granularity where latency, cost, and carbon interact and can be co-managed.

amit.more@york.ac.uk · University of York

Thank you — Questions?