CEDAR: Carbon Efficient Dynamic Allocation and Routing for Agentic LLM Inference

Abstract

LLM inference now accounts for much of operational AI compute, yet production serving stacks still optimise for a single objective, typically latency or throughput, using simple routing policies such as round robin or least connections. As AI serving scales, this focus becomes costly. Global data centres consumed 415 TWh in 2024 and are projected to reach 945 TWh by 2030, while a single LLM query can consume ten times more electricity than a conventional web search. Cost and carbon can therefore no longer be treated as secondary concerns.

Recent work has improved LLM scheduling for QoS and mixed criticality workloads [3, 1], while carbon aware systems have focused on infrastructure level elasticity and placement [2, 4]. However, these strands remain separate: existing approaches do not jointly optimise latency, cost, and schedulable carbon at the queue level for agentic LLM inference.

We address the problem of scheduling mixed criticality agentic LLM inference under competing latency, cost, and carbon constraints.

Agentic workloads make this gap more pronounced. A single coding task executed by assistants such as GitHub Copilot, Cursor, or Devin can trigger ten or more sequential LLM calls, creating persistent queue backlogs and incurring 37% KV cache recompilation overhead. These workloads are mixed criticality: interactive completions require sub second responses, while background stages such as test generation, code analysis, or summarisation can tolerate delays of several seconds. This deferability gives the scheduler room to shift less urgent work in time or across regions without violating SLOs. Current systems, however, treat requests independently and make greedy per instance decisions, leaving this opportunity largely unused.

We present CEDAR, a queue level multi objective control framework for agentic LLM inference. CEDAR jointly optimises tail latency, cloud cost, and marginal carbon emissions. Its central insight is that the queue, rather than the individual request or GPU instance, is the right control point because it is where latency pressure, resource contention, and carbon aware deferral interact.

At ingress, requests are assigned a priority tier (HIGH, MEDIUM, LOW), SLO deadline, and estimated token count. CEDAR maintains virtual queues per tier, tracking backlog, wait time, and slack for earlier overload detection than GPU utilization alone. Every 10 seconds, the controller updates routing and scaling using queue state, telemetry, and carbon signals. It manages a heterogeneous fleet across three regions on demand and spot instances with A100/H100 GPUs and optimizes using marginal carbon signals (extra emissions per decision) rather than average grid intensity.

We evaluate CEDAR via trace driven simulation that replays production LLM serving workloads with carbon intensity traces sampled at 5 minute granularity across three AWS regions. Relative to a performance only baseline, CEDAR reduces cloud cost by 26% and marginal carbon by 27%, while maintaining competitive p95 latency (0.88 s versus 0.76 s) and limiting SLO violations to 4.3%. Against round robin routing, it reduces SLO violations from 22.3% to 4.3% and lowers routing oscillation by 63%. Ablation results show that carbon aware routing, queue level slack tracking, and model right sizing each contribute to the gains.

References

[1] Kanishk Goel, Jayashree Mohan, Nipun Kwatra, Ravi Shreyas Anupindi, and Ramachandran Ramjee. Niyama: Breaking the silos of LLM inference serving. *CoRR*, abs/2503.22562, 2025.

[2] Walid A. Hanafy, Qianlin Liang, Noman Bashir, David E. Irwin, and Prashant J. Shenoy. Carbonscaler: Leveraging cloud workload elasticity for optimizing carbon-efficiency. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(3):57:1–57:28, 2023.

[3] Archit Patke, Dhemath Reddy, Saurabh Jha, Haoran Qiu, Christian Pinto, Chandra Narayanaswami, Zbigniew T. Kalbarczyk, and Ravishankar K. Iyer. Queue management for SLO-oriented large language model serving. In *Proceedings of the 15th ACM Symposium on Cloud Computing*, SoCC '24, pages 18–35. Association for Computing Machinery, 2024.

[4] Ana Radovanovic, Ross Koningstein, Ian Schneider, Bokan Chen, Alexandre Duarte, Binz Roy, Diyue Xiao, Maya Haridasan, Patrick Hung, Nick Care, Saurav Talukdar, Eric Mullen, Kendal Smith, MariEllen Cottman, and Walfredo Cirne. Carbon-aware computing for datacenters. *CoRR*, abs/2106.11750, 2021.