Towards Sustainable AI Systems

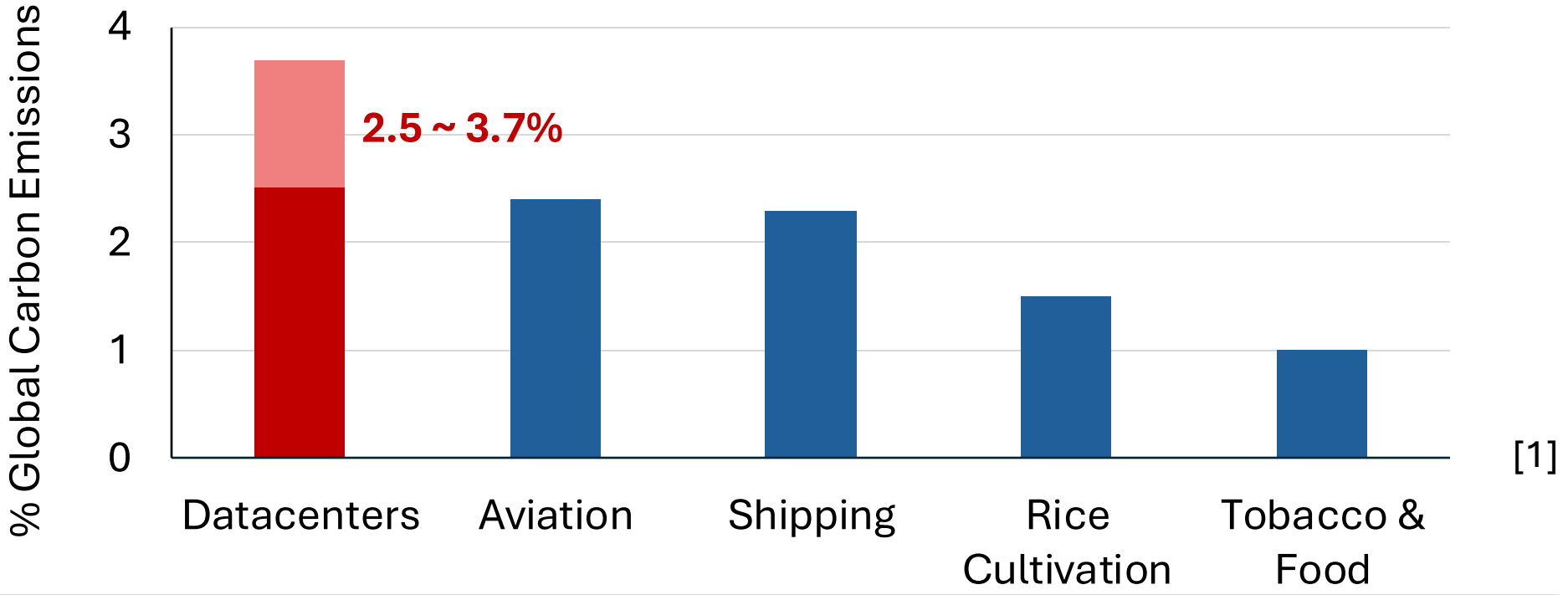
Sihang Liu

University of Waterloo

DSDS Workshop
Nov 11, 2024

UNIVERSITY OF WATERLOO

Carbon Emissions due to Computing

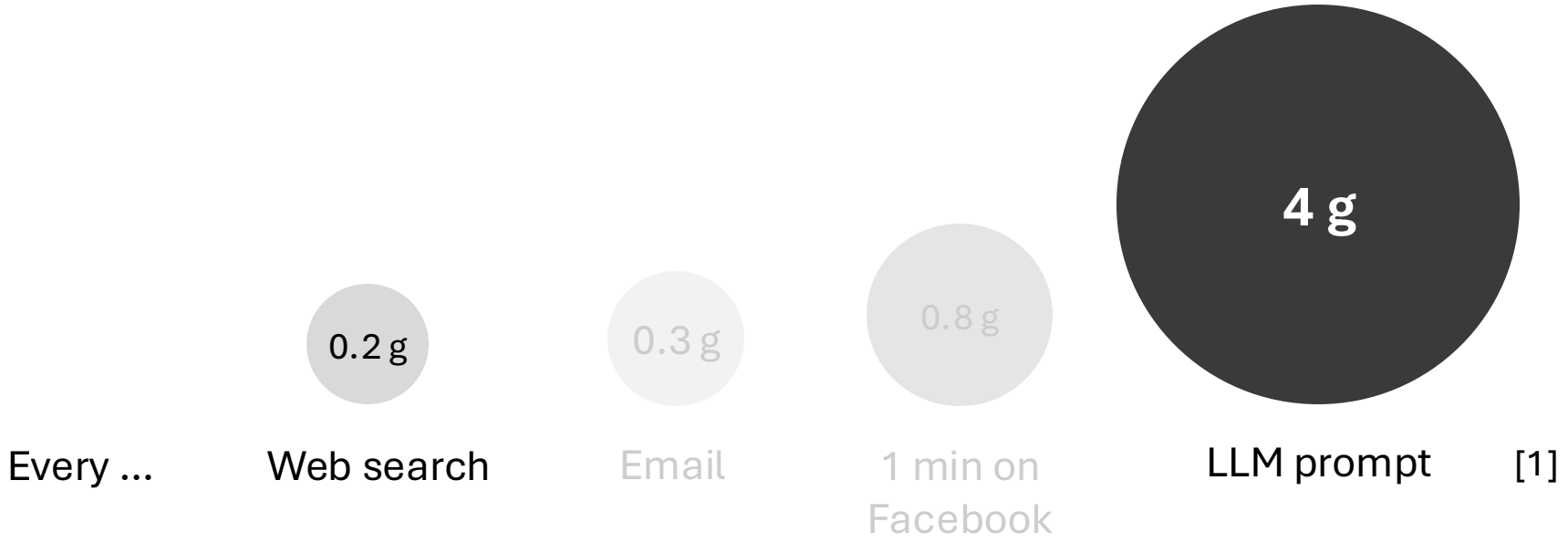


[1]

Carbon Emissions from datacenters exceeds aviation and continues to grow!

Source: Lavi. Measuring greenhouse gas emissions in data centres: the environmental impact of cloud computing. 2022

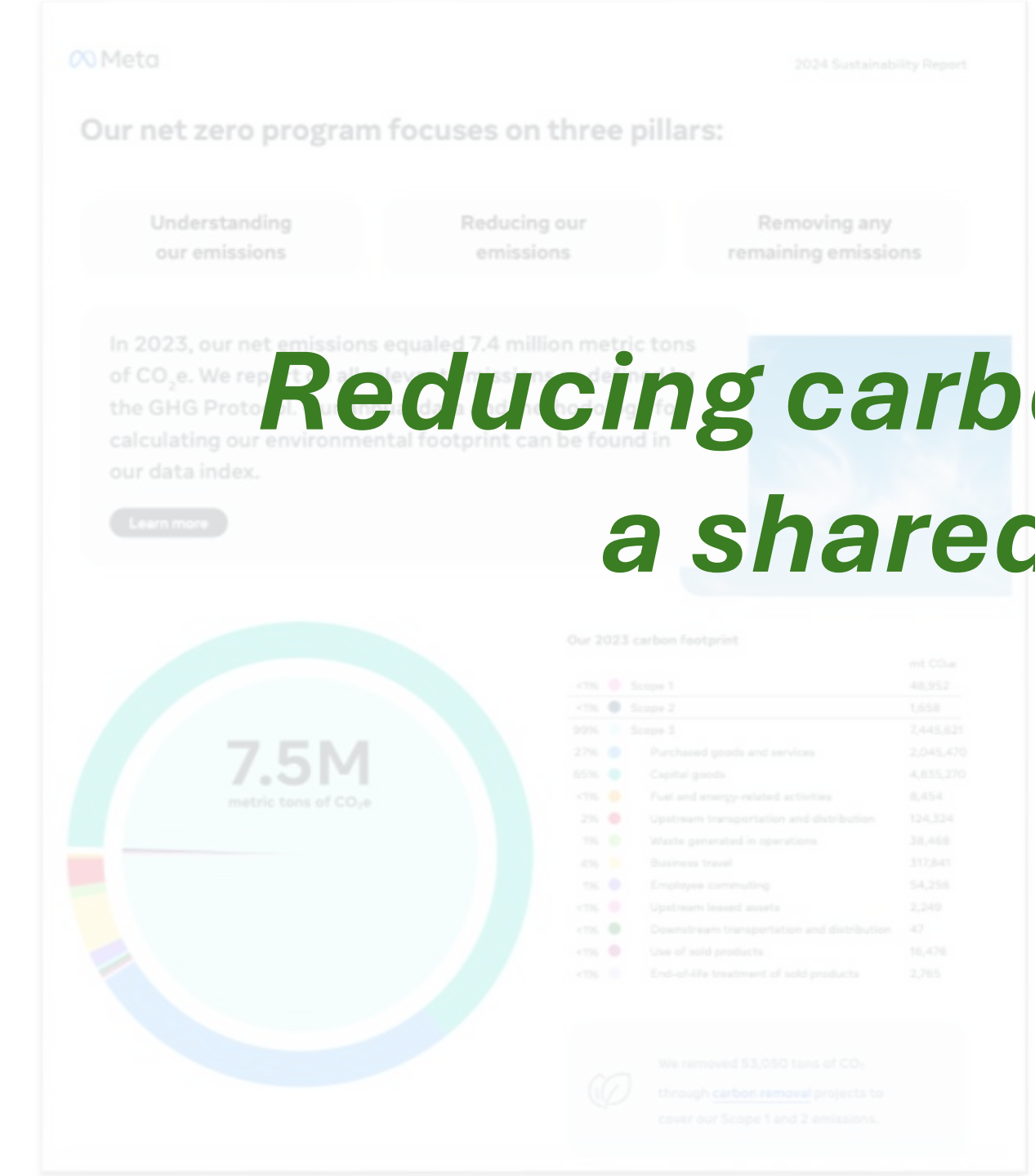
Challenges from Carbon Emissions of AI



20x more carbon emissions

[1] Wong. Gen AI's Environmental Ledger: A Closer Look at the Carbon Footprint of ChatGPT. 2024

Reducing carbon emissions is a shared mission



Carbon Emission Accounting

Operational Carbon

- From energy consumption
- Energy consumption (kWh) × Carbon Intensity (gCO_2/kWh)

Embodied Carbon

- From manufacturing process of hardware
- Longer hardware lifetime leads to lower embodied carbon

Mitigate Carbon Emissions from Cloud and AI Systems

Carbon intensity awareness

Workload and infrastructure awareness

- Uncertainty-Aware Decarbonization for Datacenters. Amy Li, Sihang Liu, and Yi Ding. HotCarbon'24
- Towards Sustainable Large Language Model Serving. Sophia Nguyen*, Beihao Zhou*, Yi Ding, and Sihang Liu. HotCarbon'24

DATA CENTERS AND INFRASTRUCTURE

Our data centers now work harder when the sun shines and wind blows

Apr 22, 2020 · 3 min read

Ana Radovanovic
Technical Lead for Carbon-Intelligent Computing

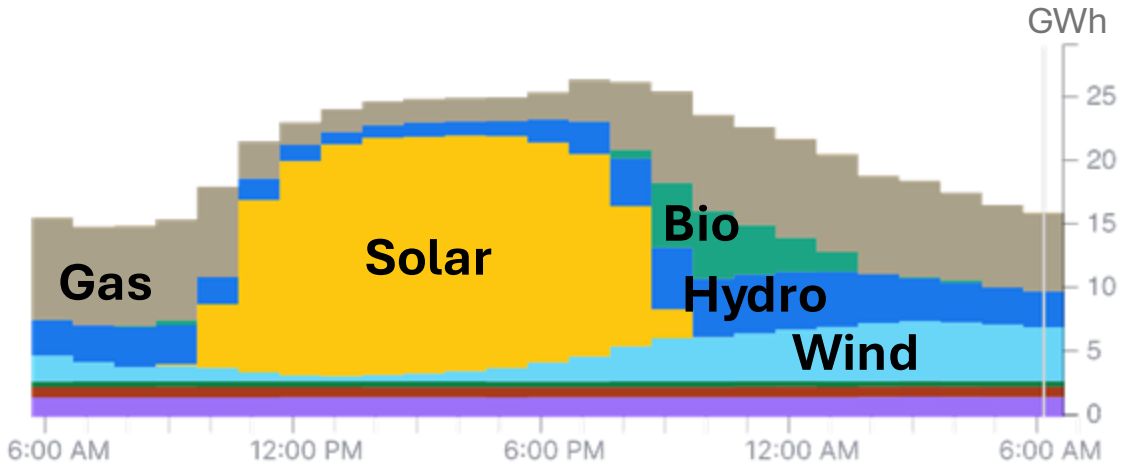
Share

Load Shifting

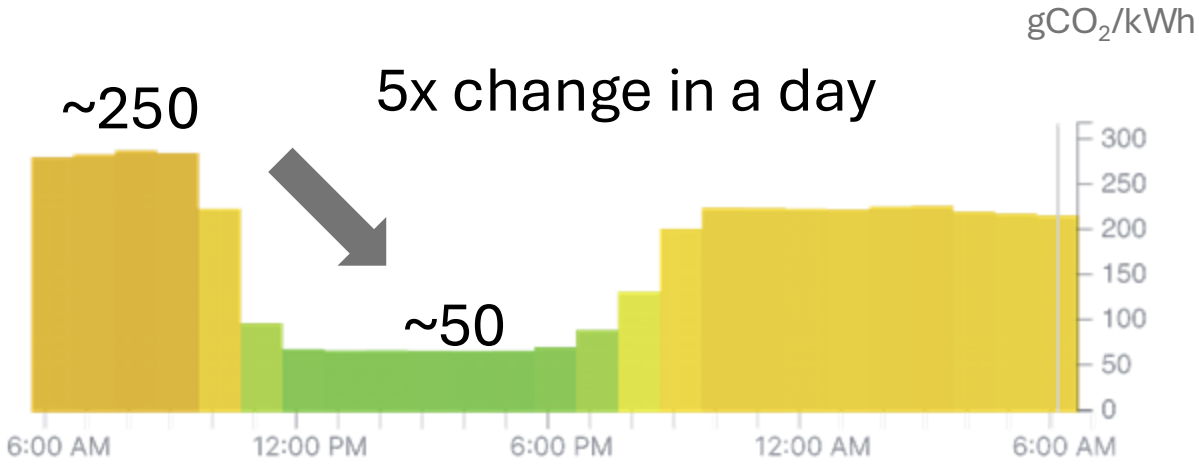
Source: Google. Our data centers now work harder when the sun shines and wind blows. 2020.

Dynamics of Carbon Intensity

- CI changes rapidly due to dynamic energy source mix
- Average CI of a region does not represent the real-time value



Energy sources of CISO (California) in a day



Change of carbon intensity in CISO

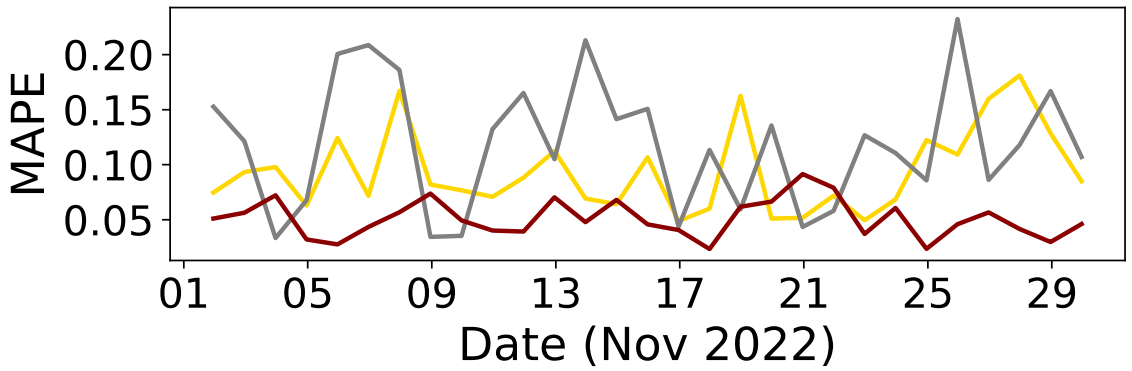
Source: electricitymaps.com

Predicting Future Carbon Intensity is Hard

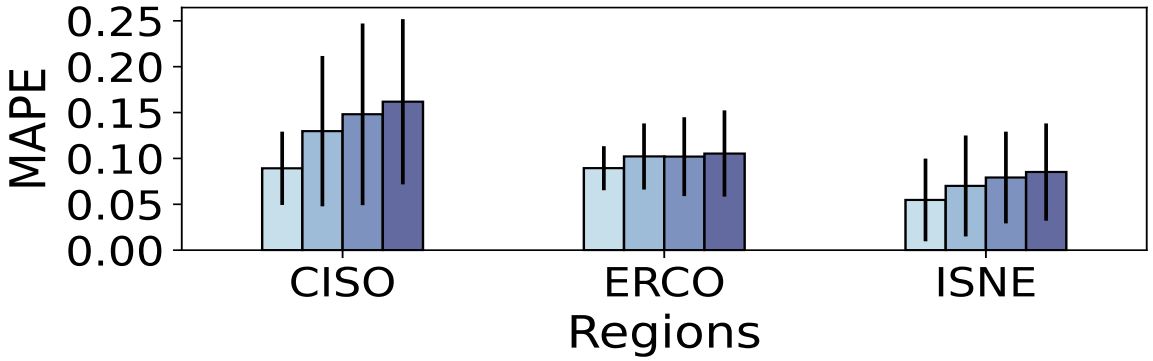
Prediction errors are due to two types of uncertainties:

Spatial and **Temporal**

Spatial uncertainty



Temporal uncertainty

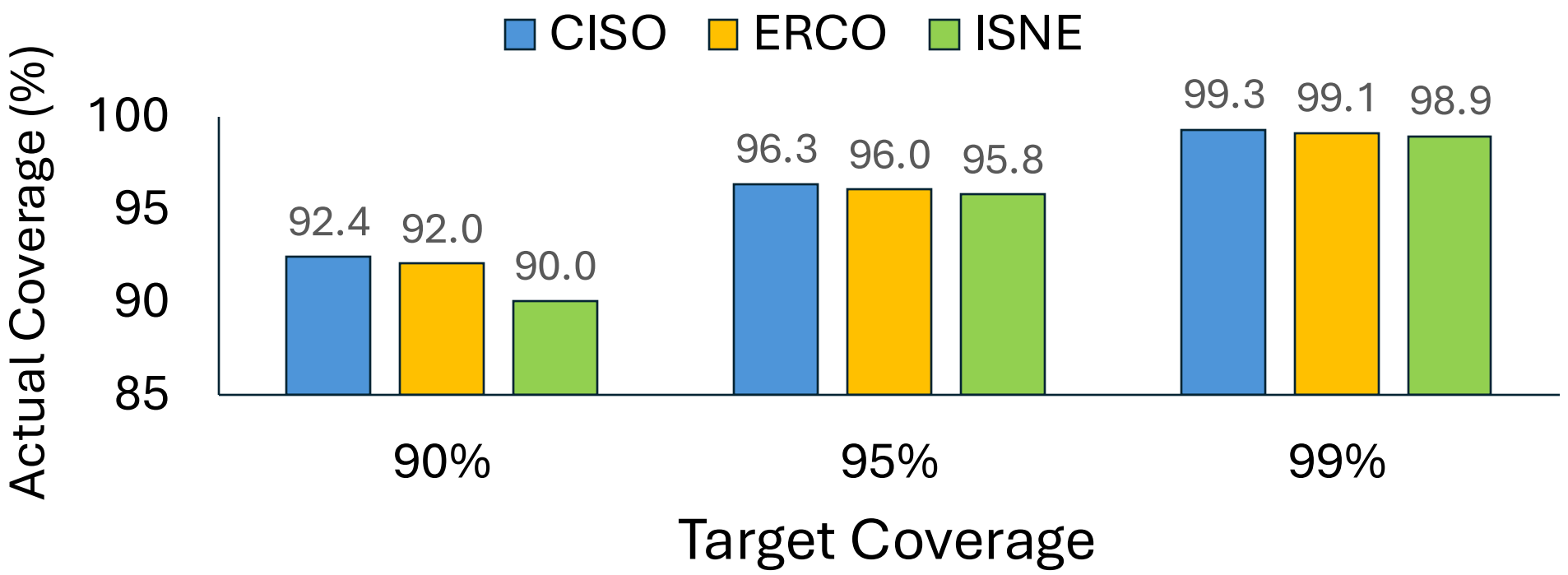


6h

Our work: Quantify uncertainty based on conformal prediction to inform more reliable load shifting.

Examples
CarbonCa

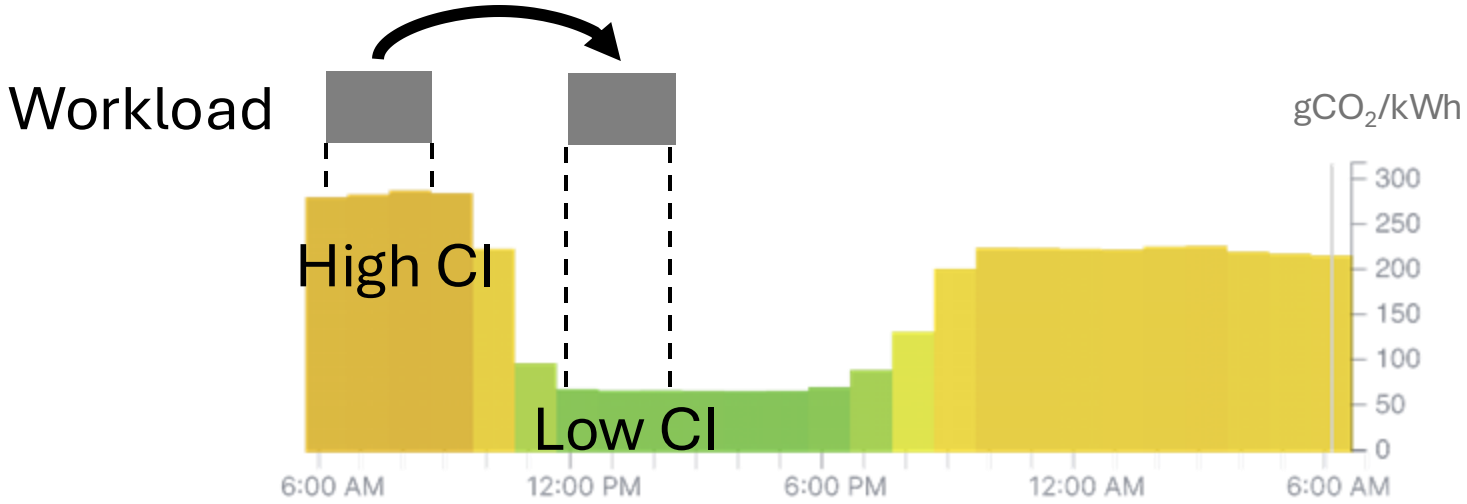
Uncertainty Quantification Results



These results highlight the efficacy of our approach in quantifying uncertainty for carbon intensity prediction.

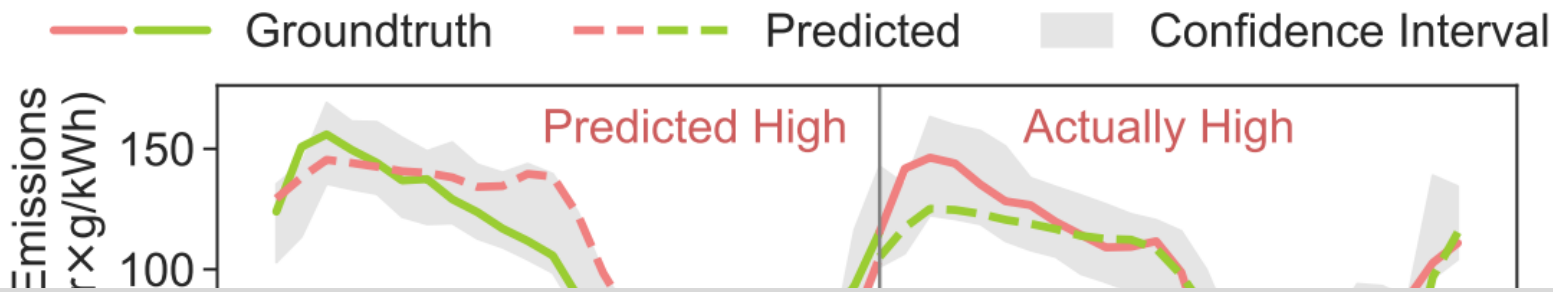
Case Study

- Simulation dataset: datacenter power traces[1]
- Load shifting policy: **Temporal shifting**
  (also known as *suspend-and-resume* or *wait-awhile*[2])
  - *Suspend* the work at higher CI
  - *Resume* the work at lower CI

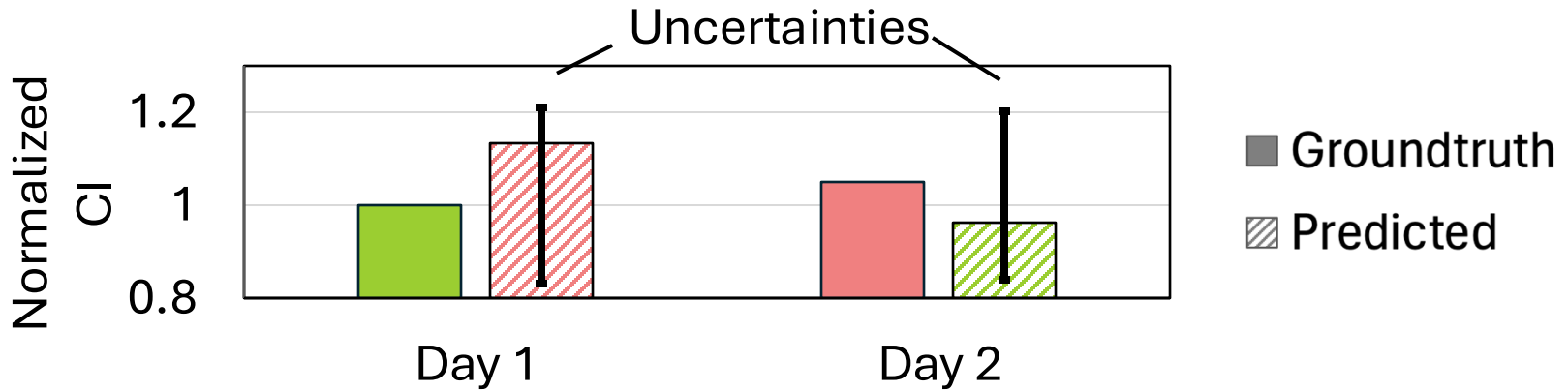


[1] Data center power oversubscription with a medium voltage power plane and priority-aware capping. ASPLOS, 2020.
[2] Let's wait awhile: how temporal workload shifting can reduce carbon emissions in the cloud. Middleware, 2021.

A Two-Day Example of Temporal Load Shifting



Load shifting should occur only when confidence is high enough



Mitigate Carbon Emissions from Cloud and AI Systems

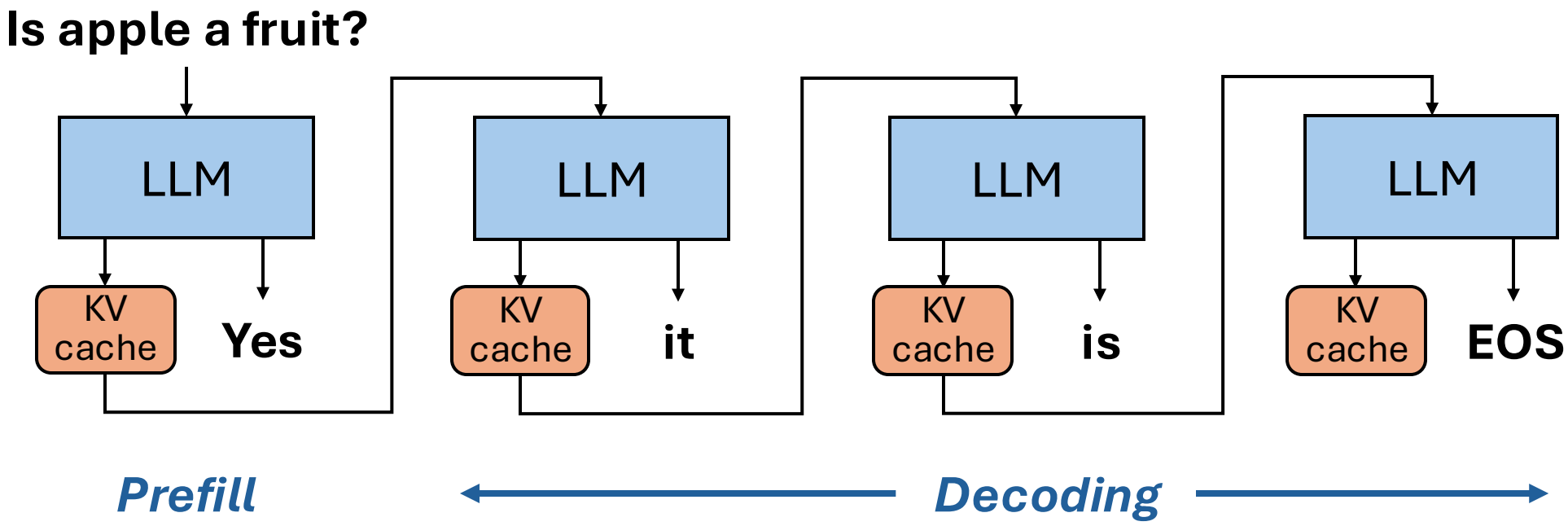
Carbon intensity awareness

Workload and infrastructure awareness

- Uncertainty-Aware Decarbonization for Datacenters. Amy Li, Sihang Liu, and Yi Ding. HotCarbon'24
- Towards Sustainable Large Language Model Serving. Sophia Nguyen*, Beihao Zhou*, Yi Ding, and Sihang Liu. HotCarbon'24

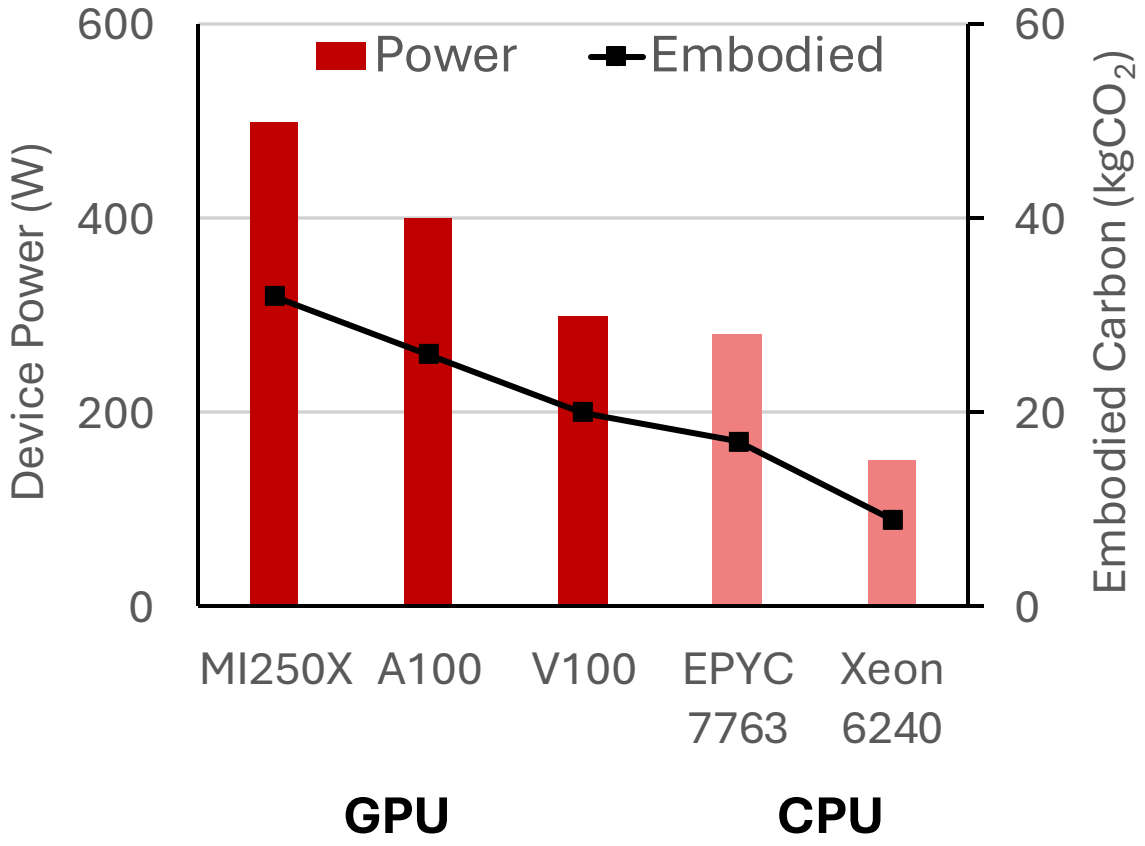
Background: Large Language Models (LLMs)

- Large language models (LLMs) are widely used in various tasks
- Inference of LLMs follow an auto-regressive pattern



Carbon Emissions of LLMs

- LLMs are compute-intensive
- Require high-end GPUs or ML accelerators
  - **High operational carbon**
  - **High embodied carbon**



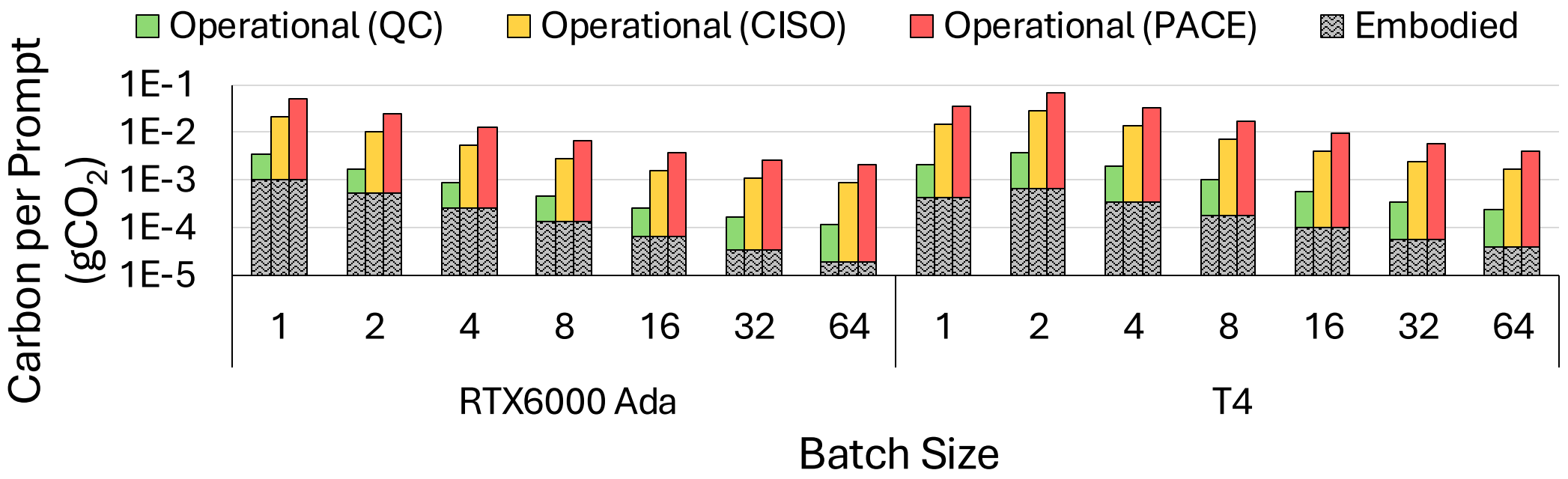
Data source: Toward Sustainable HPC: Carbon Footprint Estimation and Environmental Implications of HPC Systems. SC 2023.

Tradeoffs between Performance and Carbon

- Characterize LLM serving via token-level profiling
  - **Latency, throughput, and energy consumption**
- Carbon modeling
  - **Operational** carbon for 3 regions
    - **Low** carbon intensity: **QC** (Québec)
    - **Moderate** carbon intensity: **CISO** (California)
    - **High** carbon intensity: **PACE** (Mountain states in US)
  - **Embodied** carbon for GPUs
    - Older GPU: T4, 70W
    - Newer GPU: RTX 6000 Ada, 300W

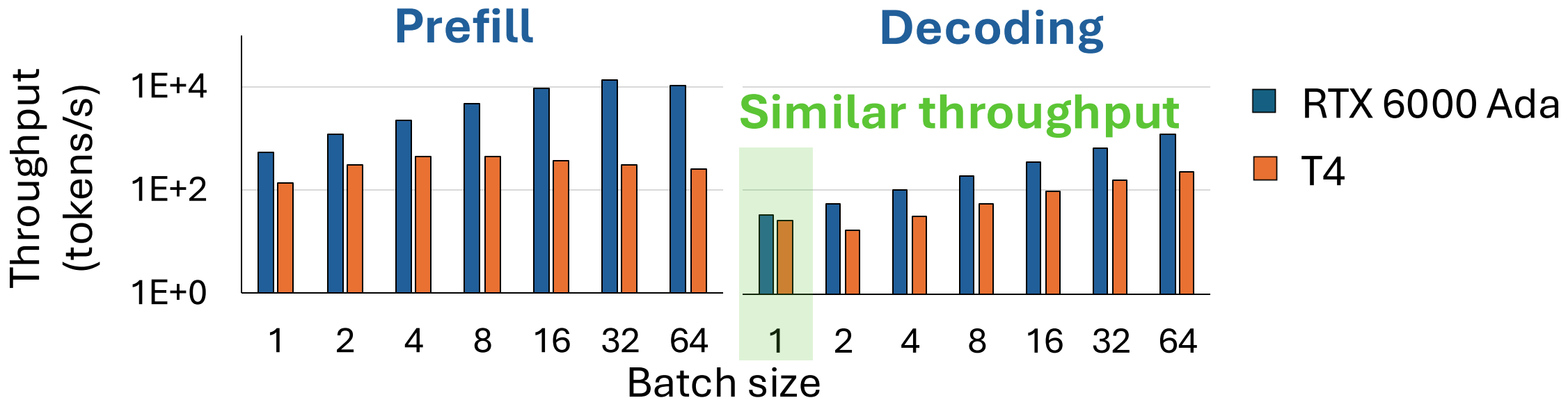
Carbon Emissions in Different Regions

- Evaluate *per-prompt* carbon emissions of 1B-parameter LLaMA with prompts from Alpaca dataset



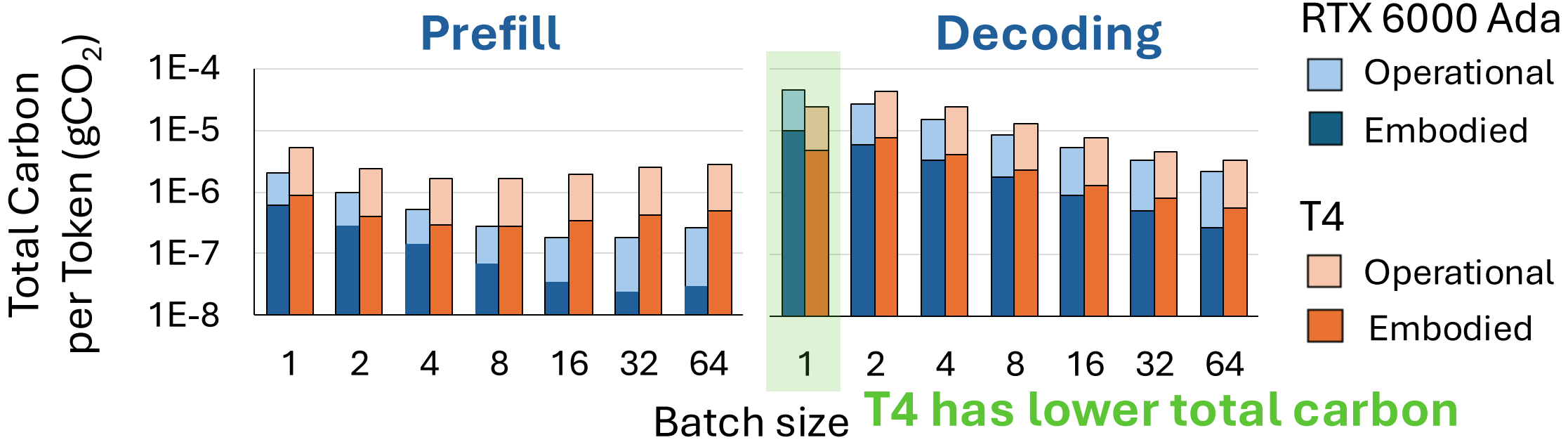
Performance vs. Carbon Emission

- Evaluate **prefill** and **decoding** phases of 1B-parameter LLaMA



Performance vs. Carbon Emission

- Evaluate **prefill** and **decoding** phases of 1B-parameter LLaMA
- Calculate total carbon emission in **QC** with **low CI**



RTX 6000 Ada is faster and has lower total carbon emissions than T4, except when batch size is 1

Takeaways

- **Uncertainty matters:**
  Carbon-aware scheduling should be done only with enough confidence
- **Region matters:**
  Older GPUs are overall less efficient but can be useful in regions with lower carbon intensities
- **Workload matters:**
  Old, lower-tier GPUs may have lower carbon emissions in less compute-intensive scenarios

Moving Forward …

- Improve both the accuracy and confidence of CI prediction with ML approaches
- Mitigate carbon emission in heterogenous AI systems
- Build an end-to-end system, from CI prediction to resource allocation, for AI applications

Towards Sustainable AI Systems

Sihang Liu

University of Waterloo

DSDS Workshop
Nov 11, 2024

UNIVERSITY OF WATERLOO