Sustainable Cloud Operations and The Role of AI

Prashant Shenoy

University of Massachusetts

March 30, 2025 @ GreenSys Workshop

Sustainability and Climate Change

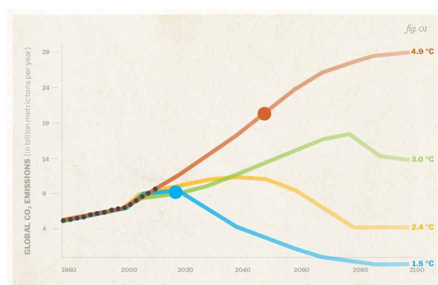
- Effects of climate change are accelerating

Climate change: Extreme weather events are 'the new norm'

By Matt McGrath
Environment correspondent

31 October

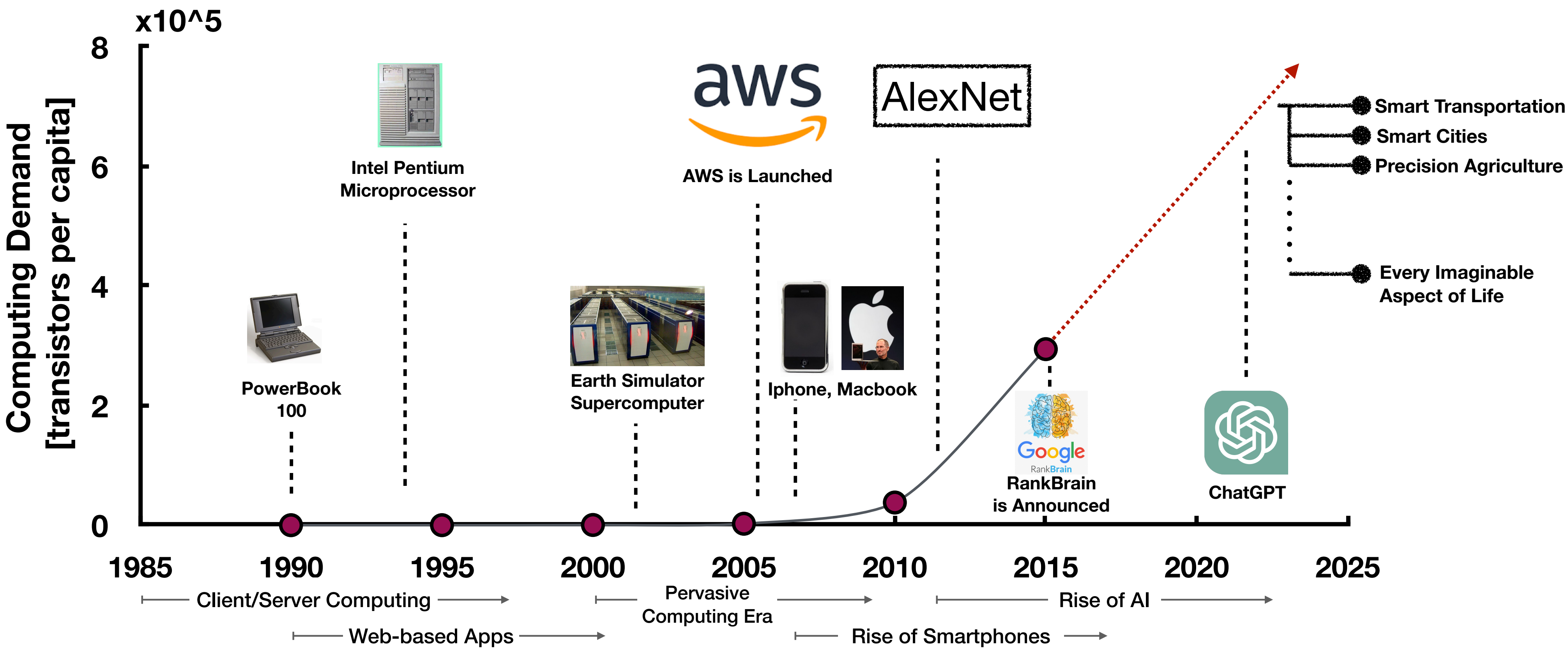
- Addressing climate change: decarbonize and reduce emissions



University of Massachusetts Amherst

Computing's Demand is Growing Exponentially

- Defining trend of our time: internet, mobile, and cloud systems



Source: "Unimaginable Output: Global Production of Transistors" - Darrin Qualman

Impact of AI Growth

- Growth driven by data-intensive and AI workloads
  - ML and deep learning workload doubling every 3.4 months
- Energy use grew more slowly due to aggressive energy/PUE optimizations
-

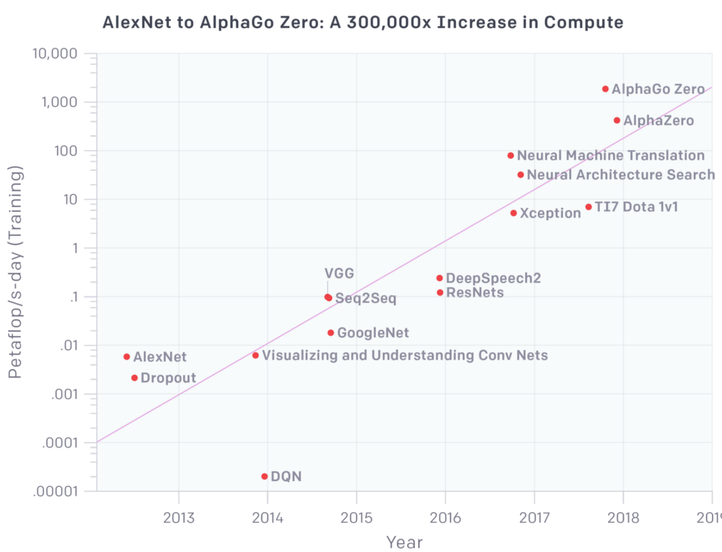


Fig courtesy: Strubell '20

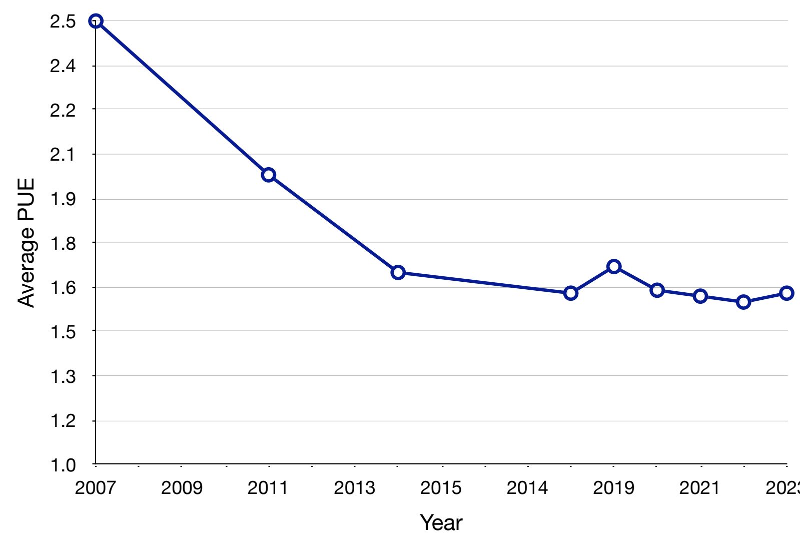


Fig courtesy: uptime institute

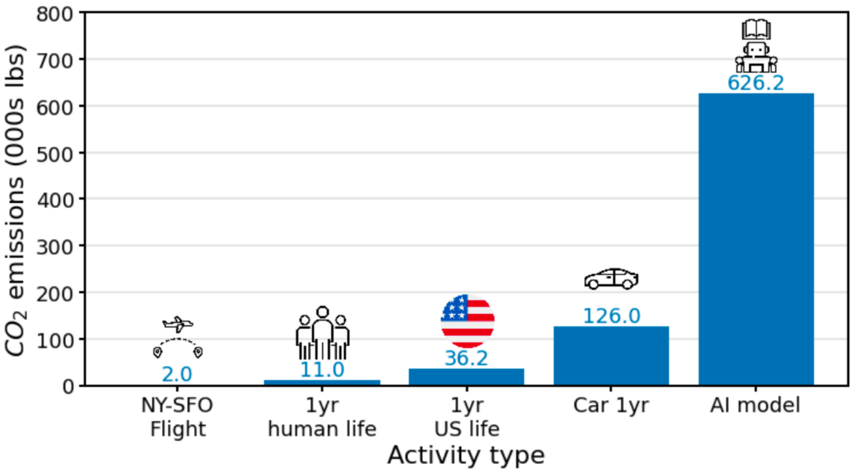
Energy efficiency vs Carbon efficiency

- **Energy efficiency**: energy consumed per unit of work done
- **Carbon efficiency**: CO_2 generated per unit of work done
- Carbon efficiency is not same as energy efficiency
  - Highly energy efficient systems can still be carbon inefficient!
- Design systems to be **both** energy- and carbon efficient

Carbon Impact of Cloud AI Workloads: How much?

- How much carbon emissions will future cloud workloads generate?

Pessimistic View



E. Strubell et. al, AAAI 2020

Optimistic View

The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink

D. Patterson et. al. IEEE Computer 2022

Both studies predated the emergence of generative AI

Research Question

- How can we use AI to decarbonize cloud infrastructure and workloads?

COMPUTING *for the*
COMMON GOOD

Talk Outline

- Motivation
- Decarbonization Basics
- Carbon First approach
- Future challenges

Decarbonizing Computing In Practice

Facebook says it has reached net zero emissions

Apple says it's now powered by 100 percent renewable energy worldwide

Carbon neutral since 2007.
Carbon free by 2030.

- **Carbon neutral**: Buy carbon offsets from energy market
  - offsets emissions
- **Net-zero** via 100% renewables: Buy renewable energy to cover electricity usage over a year
  - reduces emissions
- **24/7 matching (Carbon-free)**: Use zero-carbon energy at hourly granularity [Google'20]
  - significantly reduces emissions
- **Zero carbon**: use zero-carbon energy at "all times"

Supply-side Decarbonization Challenges

- Net-zero using 100% renewables will still generate emissions

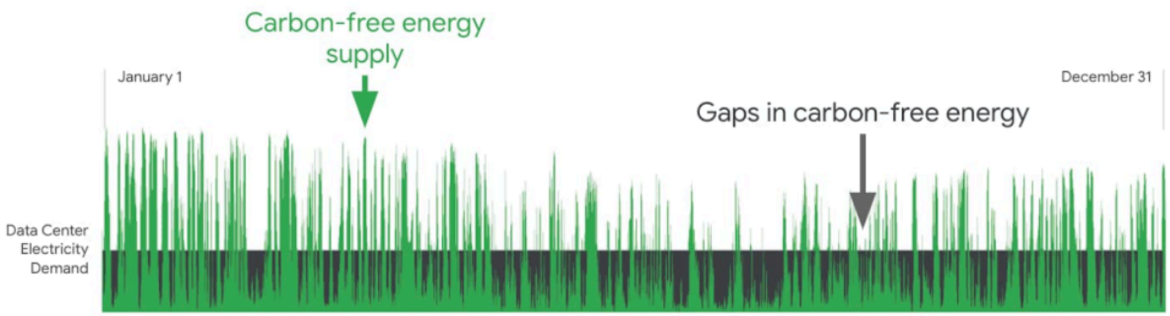


Fig courtesy: Urs Holzle

Hourly carbon-free energy performance at an example data center

- True zero carbon: needs fine time-scale matching
  - Substantially complicates energy management
  - Requires overprovisioning of renewables or zero-carbon sources such as nuclear

Decarbonization Using Demand-side Optimizations

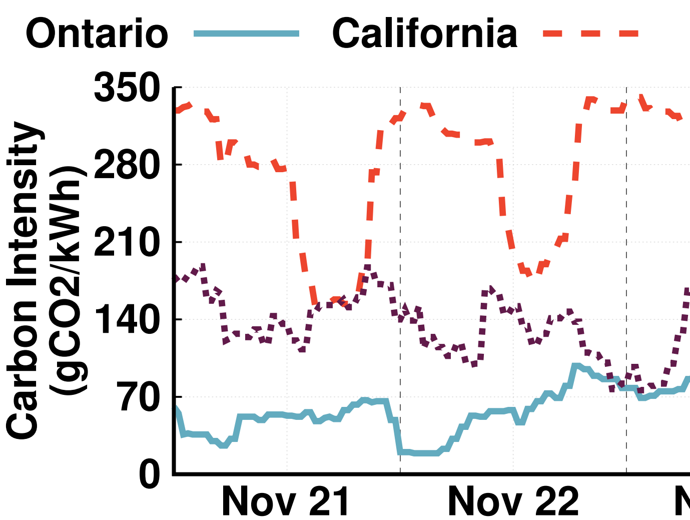
- Supply-side methods: switch to low-carbon energy sources
  - Carbon offsets, zero-carbon matching, renewable sources

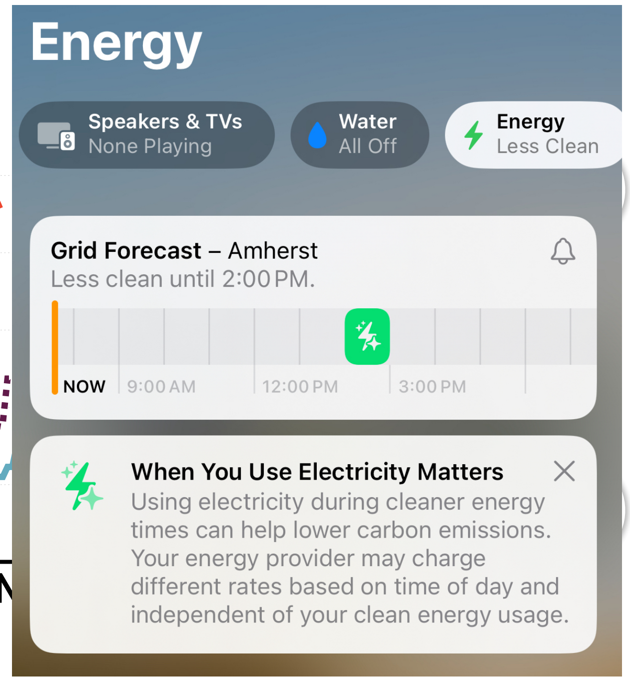
- Demand-side methods: modulate demand to reduce emissions

- Both supply and demand-side methods will be necessary to reach "true zero" emissions

- Computing workloads tend to be elastic in nature
  - Can we exploit flexibility in workload to reduce emissions?

Carbon Intensity of Electricity Varies Across Space & Time





Spatial Variations

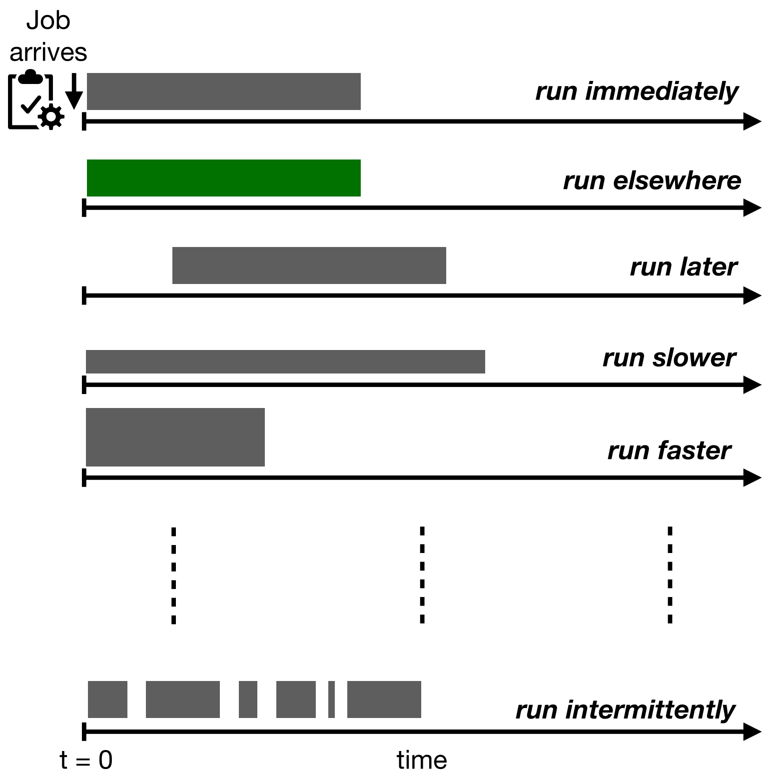
Move to the greenest data center possible

Temporal Variations

Move to a time slot with the lowest carbon emissions

Run when and where low-carbon energy is available.

Computing workloads are uniquely flexible



Driven by efforts to reduce costs, improve user experience, and scale.

University of Massachusetts Amherst

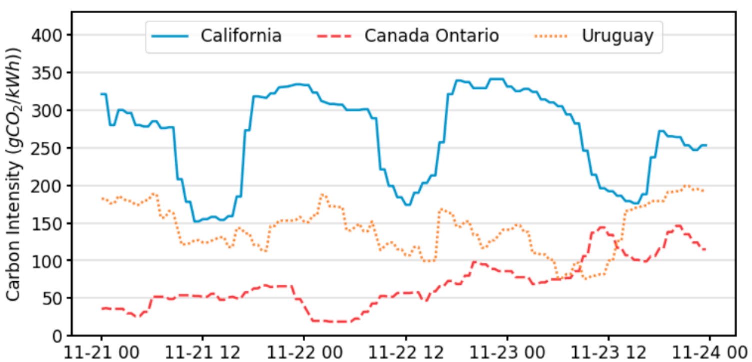
Carbon First: Decarbonizing Cloud Computing

- CarbonFirst: make carbon-efficiency first-class design concern
  - Similar to performance, reliability, …
- Key Goals:
  - Expose fine-grain energy and carbon usage to data center applications
  - Provide carbon control mechanisms to modulate carbon usage
  - Enable flexible policies to optimize the carbon usage of cloud applications
  - Promote demand-size methods that maximizes use of zero-carbon energy

University of Massachusetts Amherst

Basic Approach

- Availability of “green” electricity varies across regions and time
  - Regions with more solar/wind have lower carbon cost
- Optimize the carbon usage of elastic cloud applications
- Approach: shift cloud workloads in time & to regions with green energy



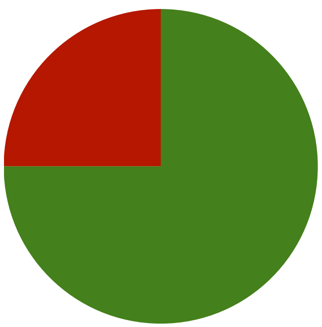
Design of green distributed cloud applications

CarbonCast: ML-driven carbon intensity forecasting.

- CI reflects the average weighted carbon intensity

$$CI = \frac{\sum (E_i \times CEF_i)}{\sum E_i}$$

Source	Coal	Natural gas	Renewables (solar, wind etc.)
CEF (g/kWh)	760	370	0



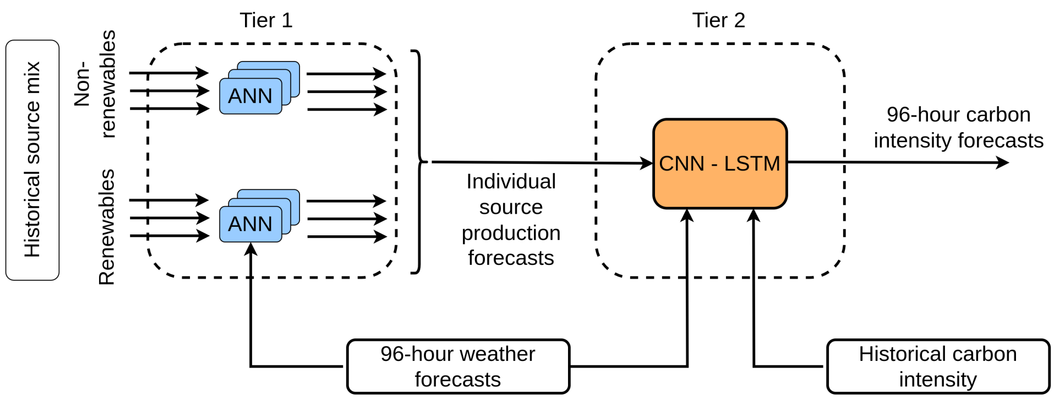
$$\begin{aligned} \text{CI} &= 760*0.25 + 0*0.75 \\ &= 190 \text{ g/kWh} \end{aligned}$$

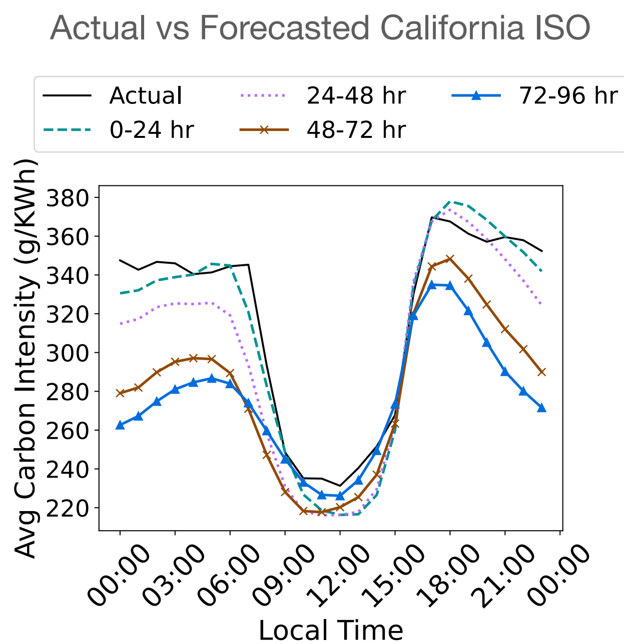
Lower CI → Greener Electricity

How can we predict future CI variations?

CarbonCast: ML-driven carbon intensity forecasting.

- Two-tier ML-based architecture





Region	MAPE
California	13.37
PJM	4.80
Germany	13.93

9.78% MAPE) on average across regions

University of Massachusetts Amherst

Carbon Control via Time Shifting

- Batch and data processing workload have time elasticity
- Wait-a-while [Wiesner 2021] - Suspend-resume approach
  - Pause computations when carbon cost is high
  - Resume computations when carbon cost is low

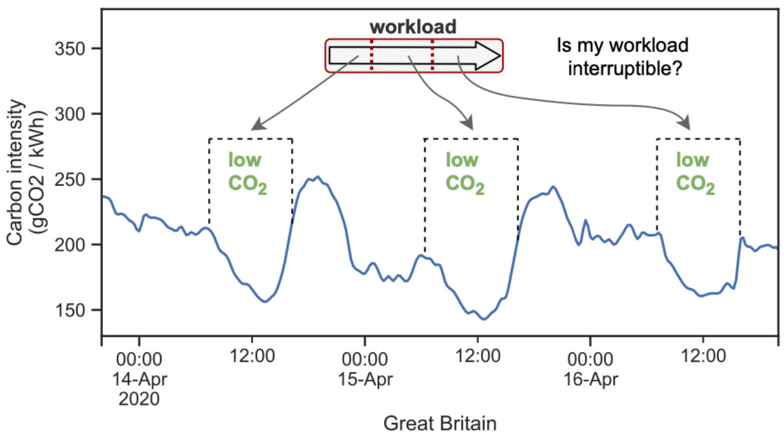
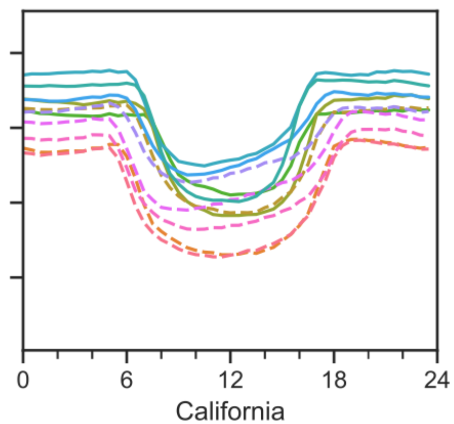


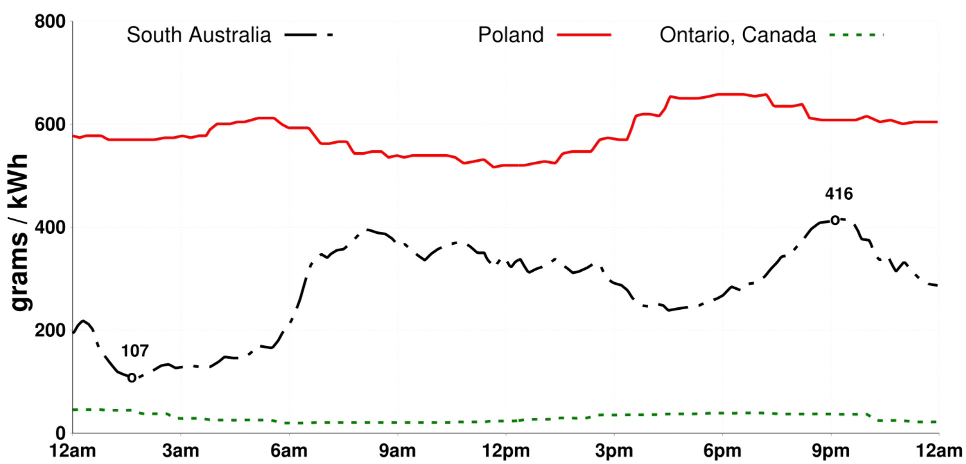
Fig courtesy: Wiesner'21



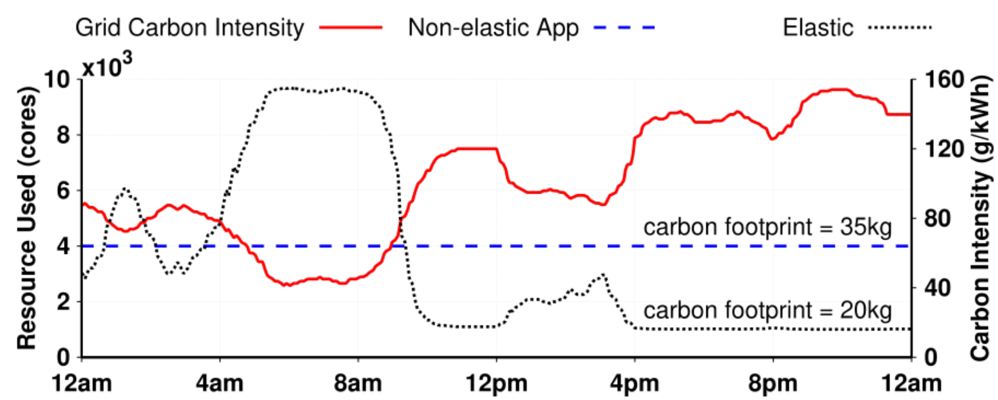
California "duck" curve

Greening Machine Learning via Continuous Scaling

- Exploit elastic nature of machine learning training
- Approach: match resources use to carbon intensity fluctuation



Carbon cost of electricity



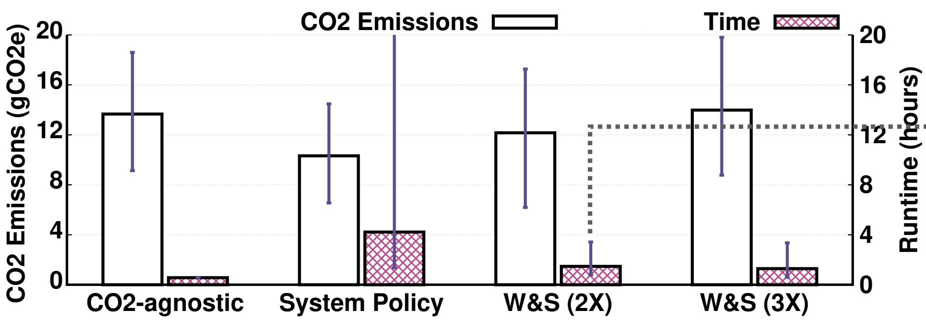
Schedule more in low carbon periods

45% carbon reduction

University of Massachusetts Amherst

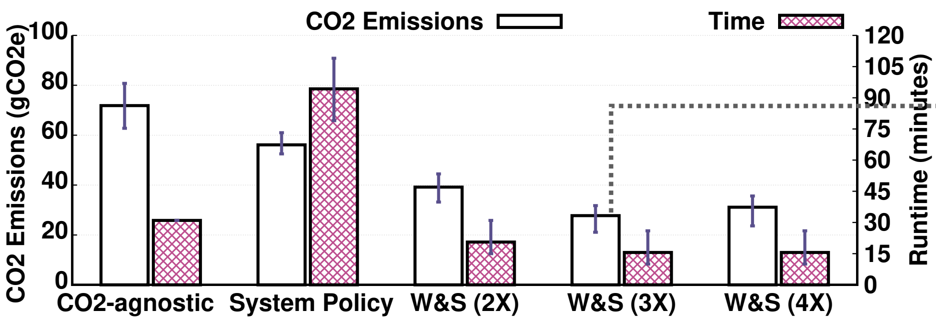
Carbon-aware Resource Scaling

- Suspend-resume increase completion time by 7X
- **Wait-and-Scale:** scale up when carbon cost is low and pause when it is high



PyTorch ML Training

Optimal Scale = 2X



BLAST

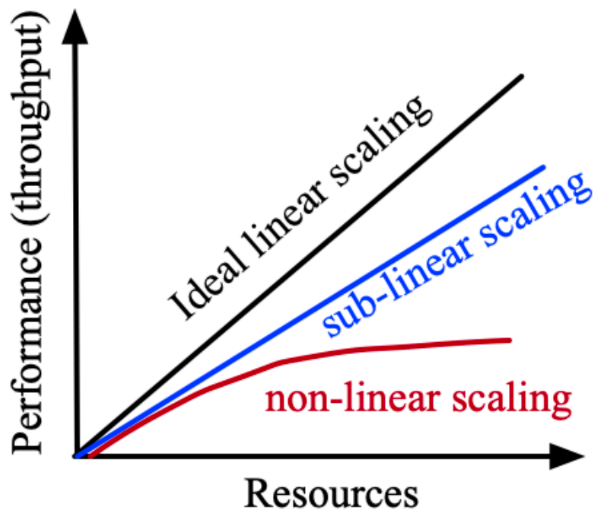
Optimal Scale = 3X

Embarrassingly parallel job.

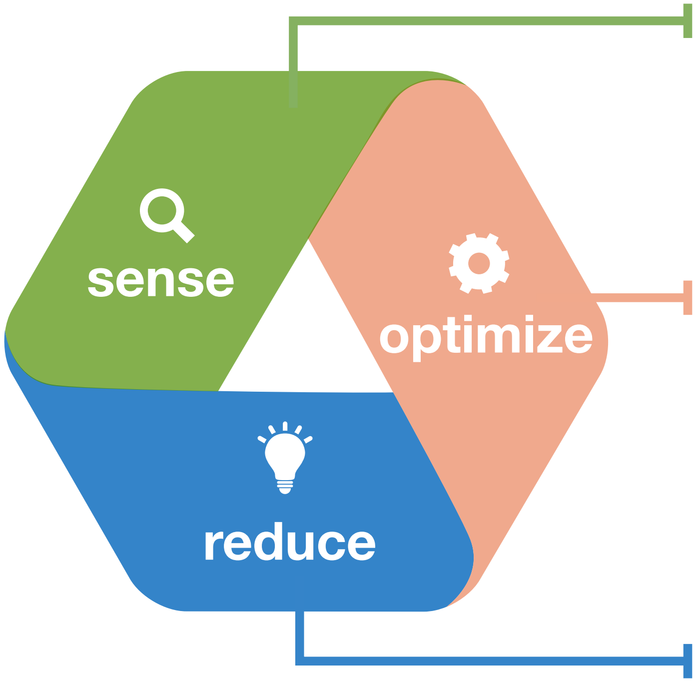
[Sigmetrics'24]

Challenges in Continuous Scaling

- Distributed cloud applications rarely scale linearly
  - Sub-linear or non-linear scaling common due to hardware/software bottlenecks
- Scaling up during low carbon periods reduces carbon efficiency!
  - Need to understand scaling behavior to implement optimal carbon-aware scaling



Decarbonizing AI



What is the carbon intensity of my electricity?
Watttime, electricityMaps, CarbonCast,

When and where should I run my training/inference?

- **CarbonScalar [Sigmetrics'24]:** Scale ML training when carbon is low.
- **DTPR [Sigmetrics'24]:** Online algorithm that considers switching costs.
- **CUFF [e-Energy'23]:** Saving power in GPU clusters.

How do I deploy these optimizations?

- **Ecovisor [ASPLOS'23]:** A carbon-aware hypervisor.
- **GAIA [ASPLOS'24]:** A carbon-aware cloud scheduler.

Concluding Remarks

- Computing systems need to become sustainable
  - AI-based approaches hold promise

- Exploit elasticity in computing workloads to reduce carbon footprint

- Significant challenges remain and will to be addressed in coming decades

- New project: NSF CoDec — Computational Decarbonization of Societal Infrastructure

Thank you

- Questions?

- http://codecexp.us and http://lass.cs.umass.edu

- Acknowledgements: A. Lechowicz, Q. Liang, W. Hanafy, D. Maji, A. Souza, N. Bashir, D. Irwin