UMassAmherst | Manning College of Information & Computer Sciences

A Hitchhiker's Guide to Sustainable Computing

Walid A. Hanafy

March 2025

A little bit about my work

Energy-Efficient Inf. [WEEE'21]

Ibis [TAAS'23]

Understanding [IWQoS'23]

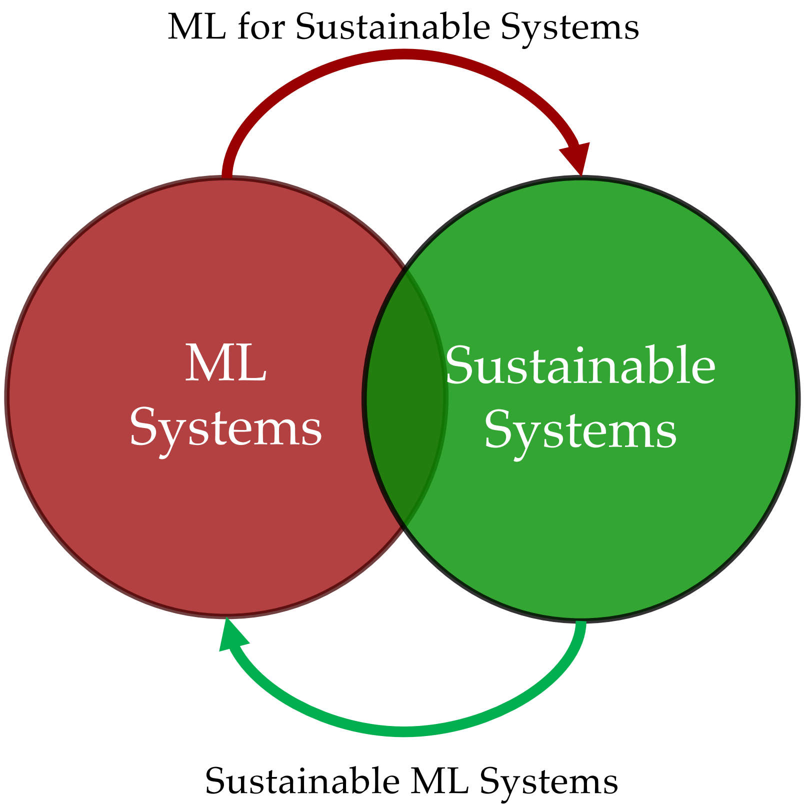
Delen [IoTDI '23]

Resilient ML [MILCOM'23]

ETF [SEC'23]

ML Pipelines [MILCOM'24]

ACIES-OS [ICCCN'24]



Efficiency Wars [HotCarbon'23]

Ecovisor [ASPLOS'23]

CarbonScaler [SIGMETRICS'24]

GoingGreen [ASPLOS'24]

LACS [e-Energy'24]

FTL [HotCarbon'24]

CDN-Shifter [SoCC'24]

GreenThrift [JCSS'25]

CarbonEdge [UnderReview]

Untangling [UnderReview]

CarbonFlex [UnderReview]

First Author

What are Sustainable Systems?

Supply Chain

Devices and systems with low embodied footprint.

Energy Efficiency

Optimizing hardware and software to consume less power.

Renewable Energy

Utilizing green (low-carbon) energy sources.

End-of-Life

Recycling and responsible disposal of electronic waste.

What are Sustainable Systems?

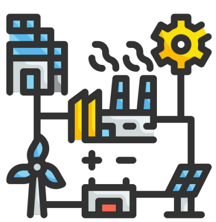
Supply Chain
Devices and systems with low embodied footprint.

Energy Efficiency
Optimizing hardware and software to consume less power.

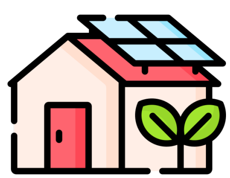
Renewable Energy
Utilizing green (low-carbon) energy sources.

End-of-Life
Recycling and responsible disposal of electronic waste.

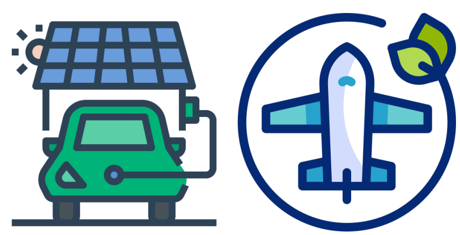
Grid



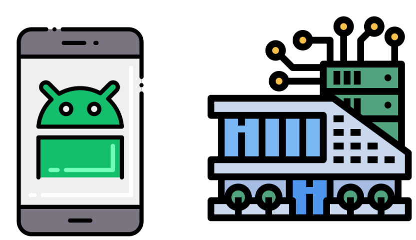
Buildings



Transport

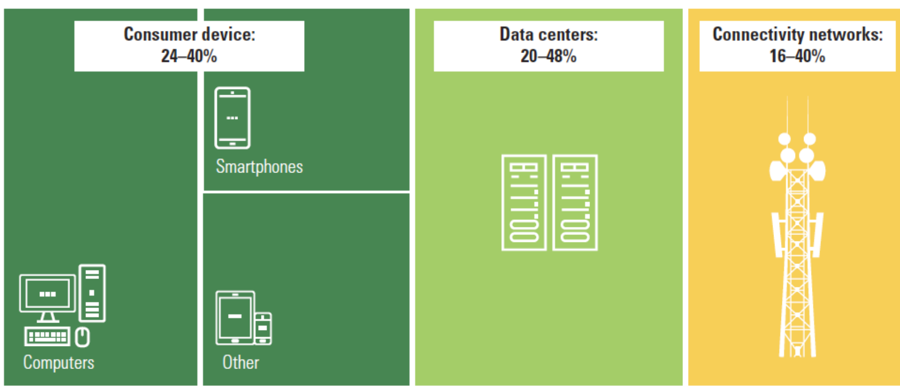


Compute



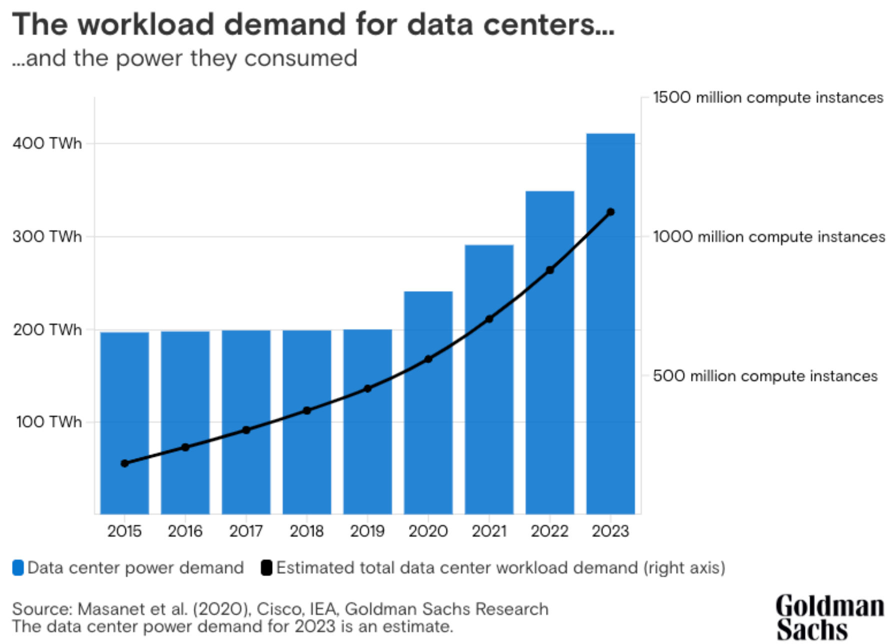
Carbon Emissions of Data Centers

ICT is responsible for **1.5 - 4%** of Global Carbon Emissions, may reach **6-14%** by 2040.



Source: Adapted from WIK-Consult and Ramboll (2021) to include estimates by Minges, Mudgal, and Decoster (forthcoming) based on analysis of reported emissions by more than 150 international digital companies.

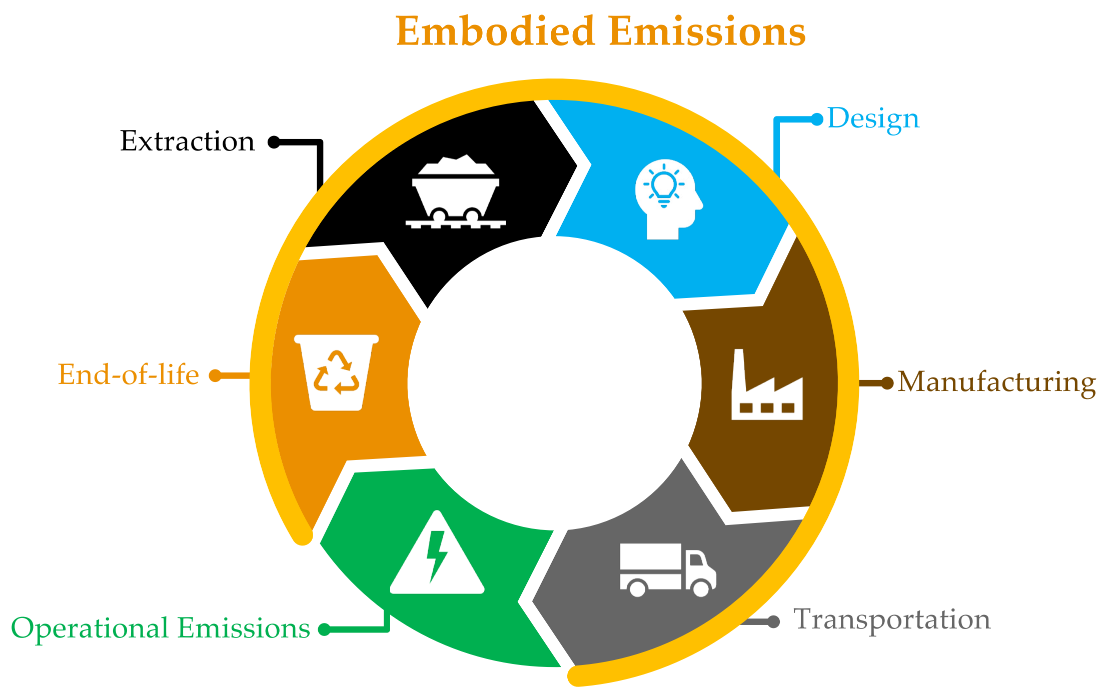
Worldbank. Green Digital Transformation. 2024



AI is poised to drive 160% increase in data center power demand.

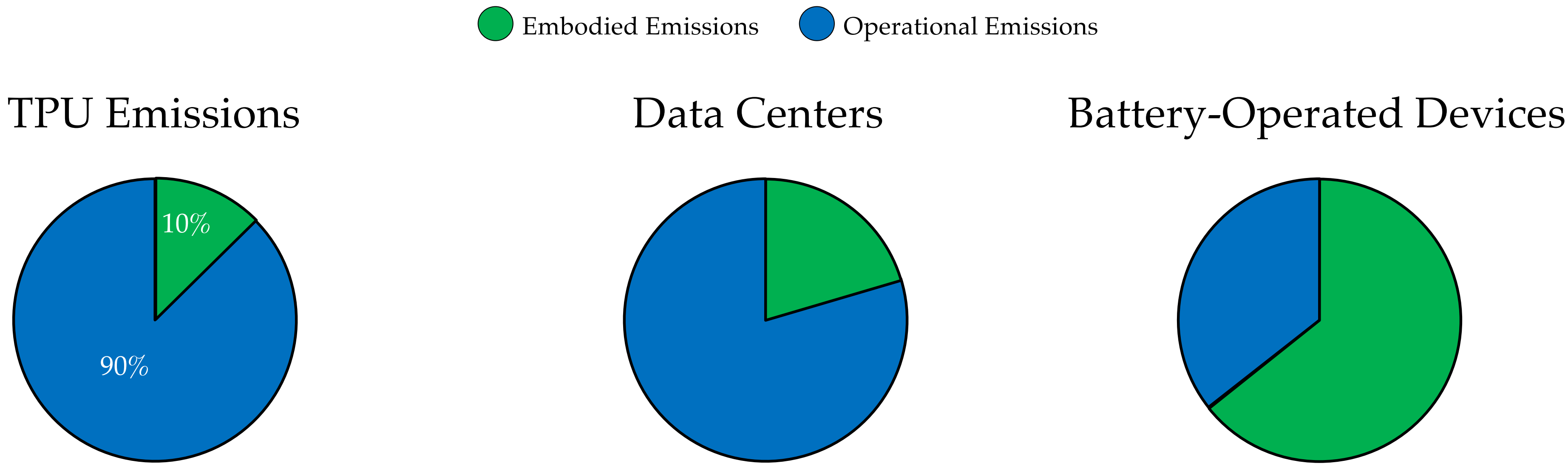
UMassAmherst | Manning College of Information & Computer Sciences

Carbon Emissions of Data Centers



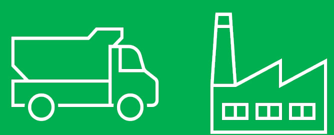
Carbon Emissions of Data Centers

Carbon Emissions of Data Centers



Optimizing Carbon Emissions of Data Centers

Embodied Emissions



- The relationship between design and embodied emissions
- Optimize Design and Manufacturing
  - ACT [ISCA'22], Focal [ASPLOS'24]
- Extending Lifetime
  - Junkyard Computing [ASPLOS'23]

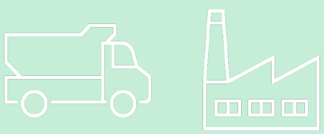
Operational Emissions

$$\uparrow \textbf{Carbon Efficiency} = \frac{\uparrow \textbf{Energy Efficiency}\,(Cycles/kWh)}{\downarrow \textbf{Carbon Intensity}\,(g.CO_2eq)}$$

- Energy Efficiency = Algorithmic Efficiency (Less Cycles) + Power Efficiency (Cycles / Watt)
- Carbon Intensity = Renewable Sources and Carbon-aware Scheduling

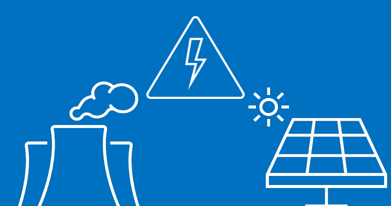
Optimizing Carbon Emissions of Data Centers

Embodied Emissions



- The relationship between design and embodied emissions
- Optimize Design and Manufacturing
  - ACT [ISCA'22], Focal [ASPLOS'24]
- Extending Lifetime
  - Junkyard Computing [ASPLOS'23]

Operational Emissions



$$\uparrow \textbf{\textit{Carbon Efficiency}} = \frac{\uparrow \textbf{\textit{Energy Efficiency}}\,(Cycles/kWh)}{\downarrow \textbf{\textit{Carbon Intensity}}\,(g.CO_2eq)}$$

- Energy Efficiency = Algorithmic Efficiency (Less Cycles) + Power Efficiency (Cycles / Watt)
- Carbon Intensity = Renewable Sources and Carbon-aware Scheduling

Algorithmic Efficiency can be further improved, but has limits

Industry has strong incentive to improve the **algorithmic efficiency**

Recent focus on ML training and Crypto-mining

[bounded]

[unbounded]

Datacenter capacity **increased by 6X** from 2010-2018

Crypto-mining and ML demand is **outpacing Moore's law**

Industry has strong incentive to **maintain and accelerate growth**

$$\textbf{Carbon Footprint} = \frac{\text{Cycles per Unit Work} \times \text{Total Units of Work}}{\text{Computing's Energy Efficiency} \times \text{Energy's Carbon Efficiency}}$$

[Koomey's Law: Energy efficiency doubles every 1.5-2.6 years**]** transition to cloud, dedicated hardware

[Laundar's Principle: Theoretical limit to be reached in 2050, practical sooner**]**

[Jevon's Paradox: Historically, gains in efficiency have not reduced demand**]**

[bounded]

[unbounded]

Zero-carbon energy means **carbon efficiency can be infinite**

Industry has helped subsidize zero-carbon energy

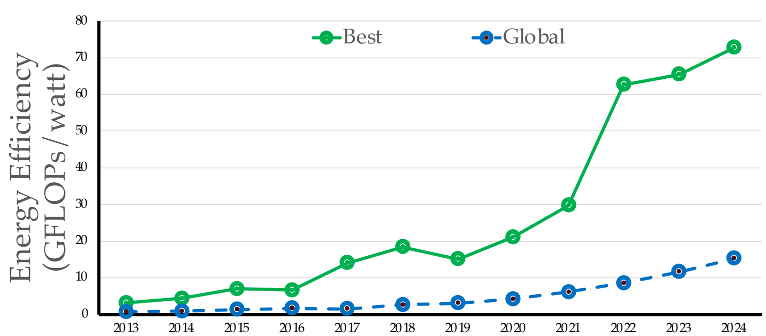
x

Optimizing Energy Efficiency

Servers (CPUs/GPUs)

- Data centers host 1000s of CPUs and GPUs that consume lots of energy.

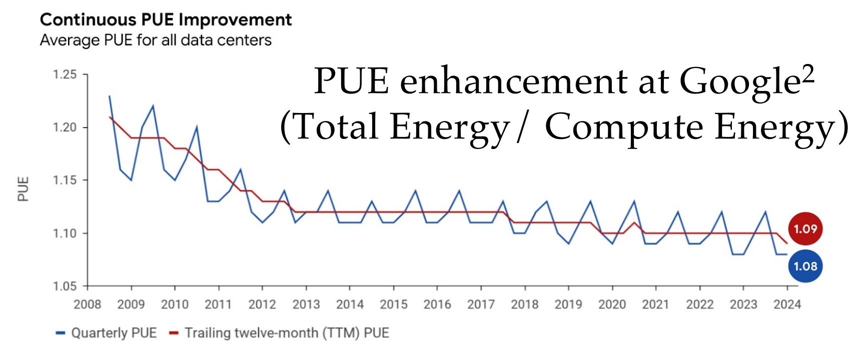
Better H/W (Gains in Energy Efficiency)[1]



Power and Resource Management (20+ Years of research)

Cooling

- Cooling: Avoid hardware failure and improve server performance.
- Cooling Innovations:
  - Raise Floors – Open Air Cooling – Liquid Cooling

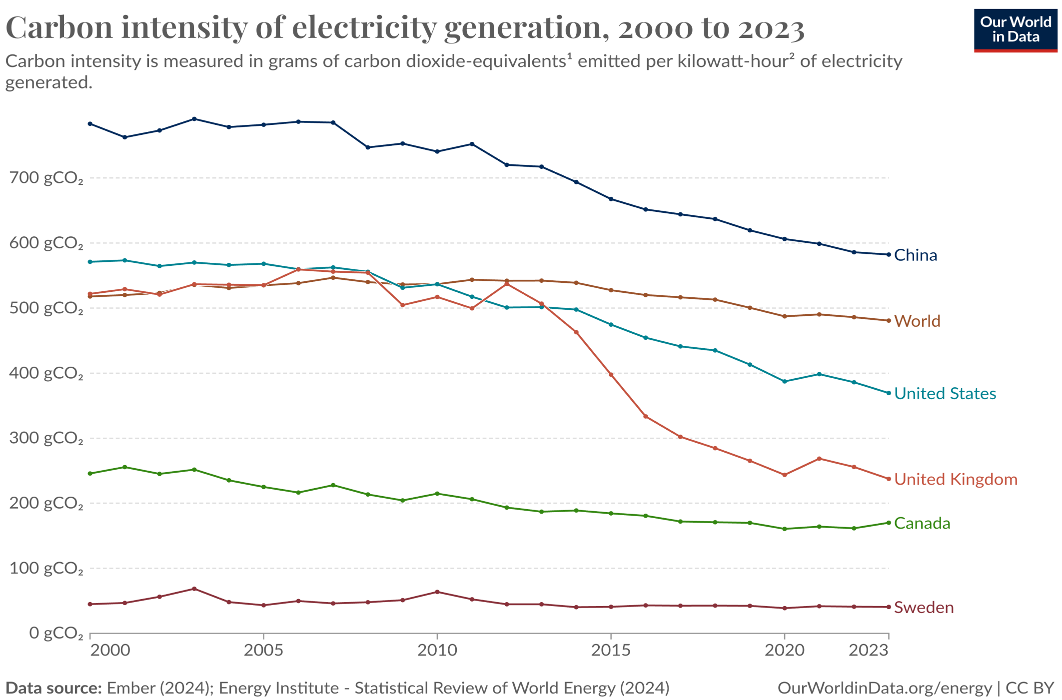


[1] https://top500.org/lists/green500/

[2] https://www.google.com/about/datacenters/efficiency/

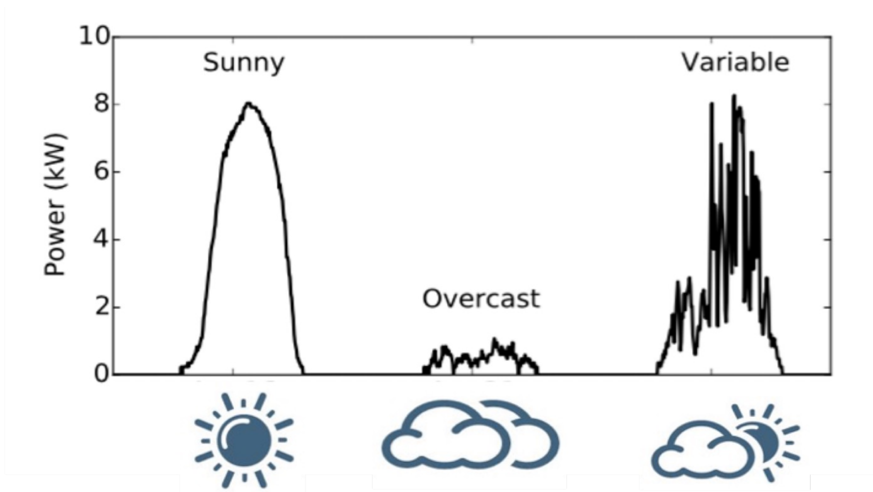
Optimizing Carbon Intensity: Renewables

- Replace fossil fuels with renewable sources.
- Adding Renewable is cost and carbon-efficient.
- Carbon Intensity is continuously decreasing.



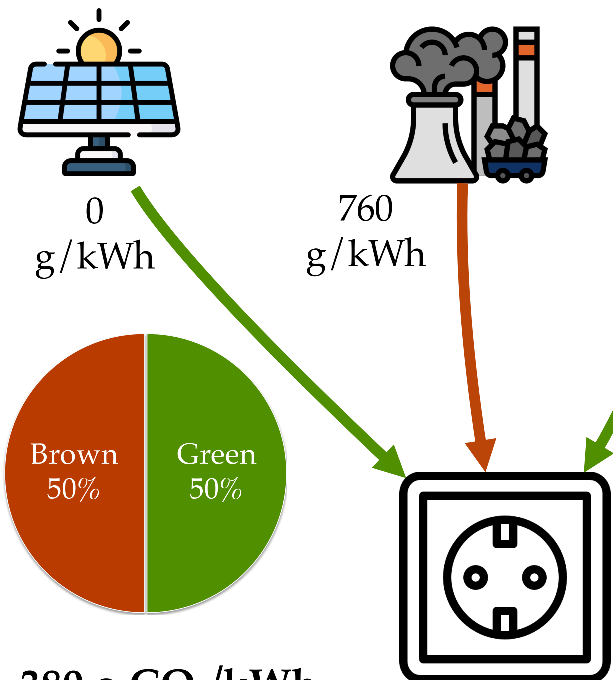
Optimizing Carbon Efficiency: Load Shifting

- Renewables are highly intermittent.
  - Solar is available in the daytime.
  - Solar affected by weather



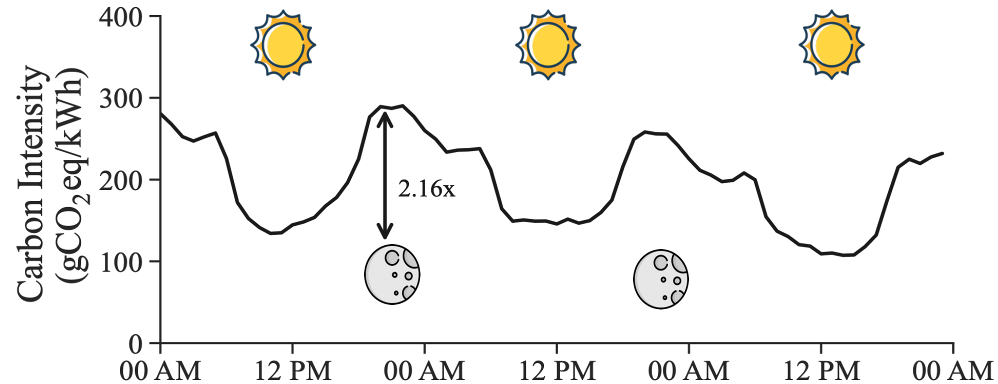
Solution: Load Shifting (Demand-Response)

Energy's Carbon Intensity ($g.CO_2/kWh$)



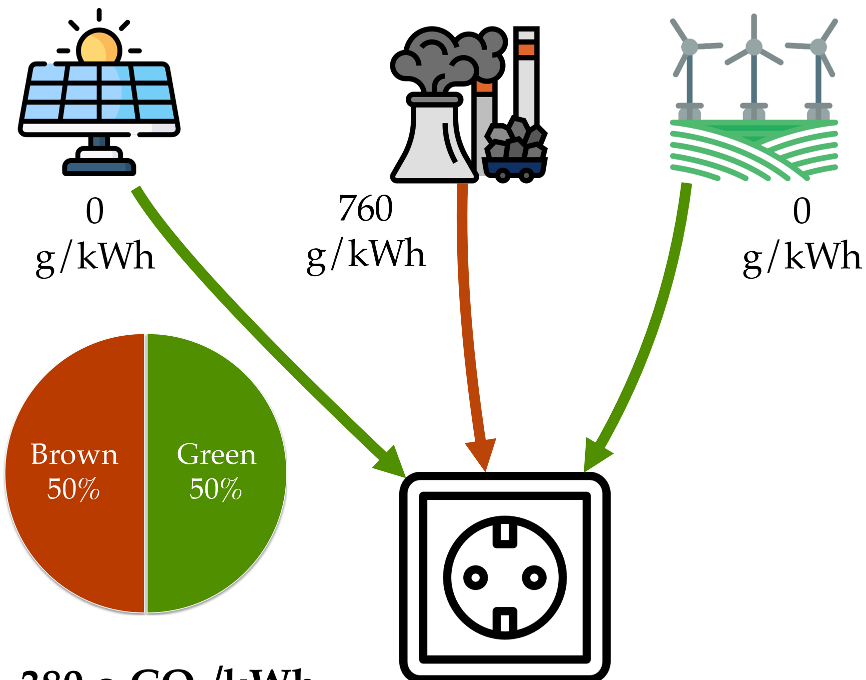
380 $g.CO_2/kWh$

Energy Demand and Supply mix change over time.



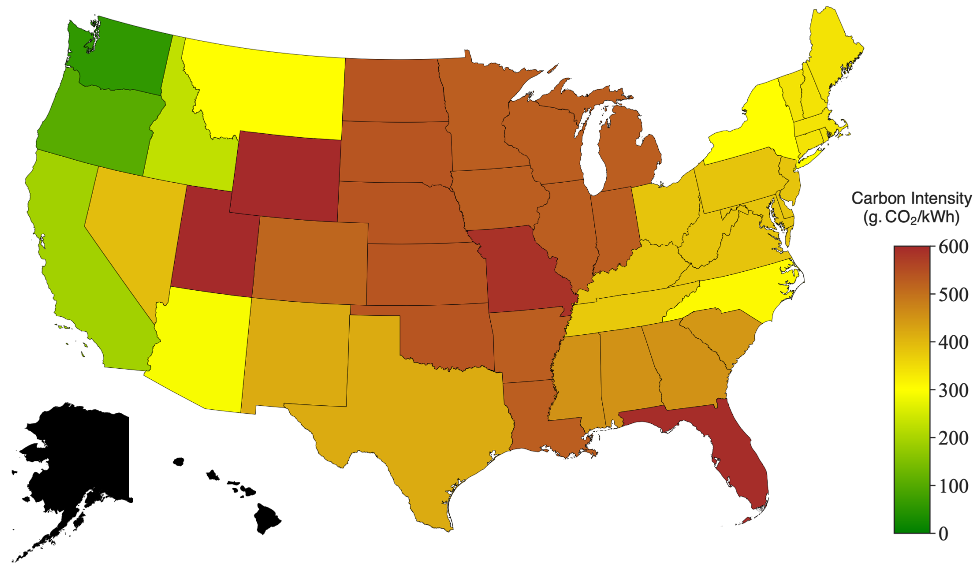
src: electricitymaps.com

Energy’s Carbon Intensity ($g.CO_2/kWh$)



380 $g.CO_2/kWh$

Energy Demand and Supply mix changes across space.

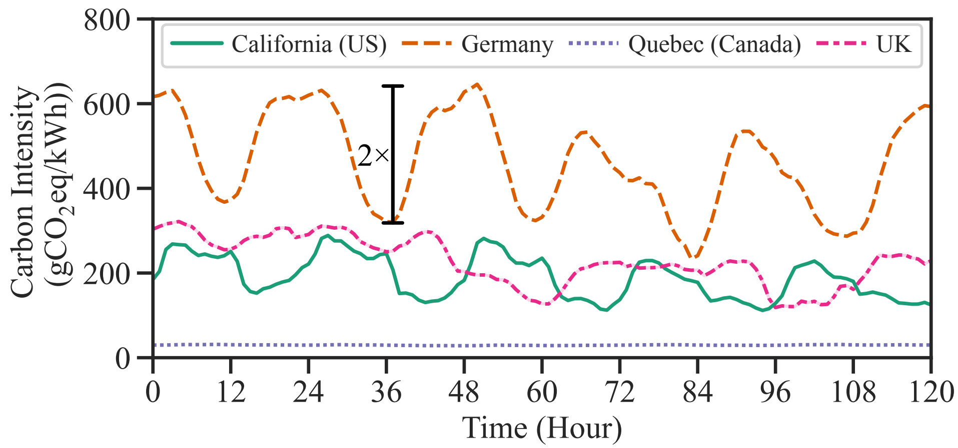


Temporal and Spatial variability underscores the need for carbon-aware resource management.

src: electricitymaps.com

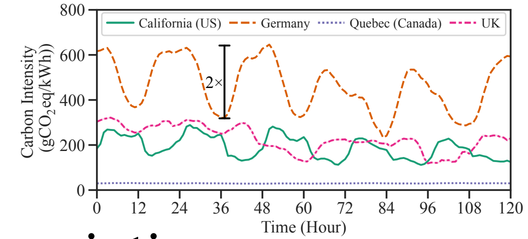
Energy's Carbon Intensity (g.CO_2/kWh)

- Carbon Intensity varies temporally by 2×.
- Carbon Intensity differs by ~600 g.CO_2/kWh.

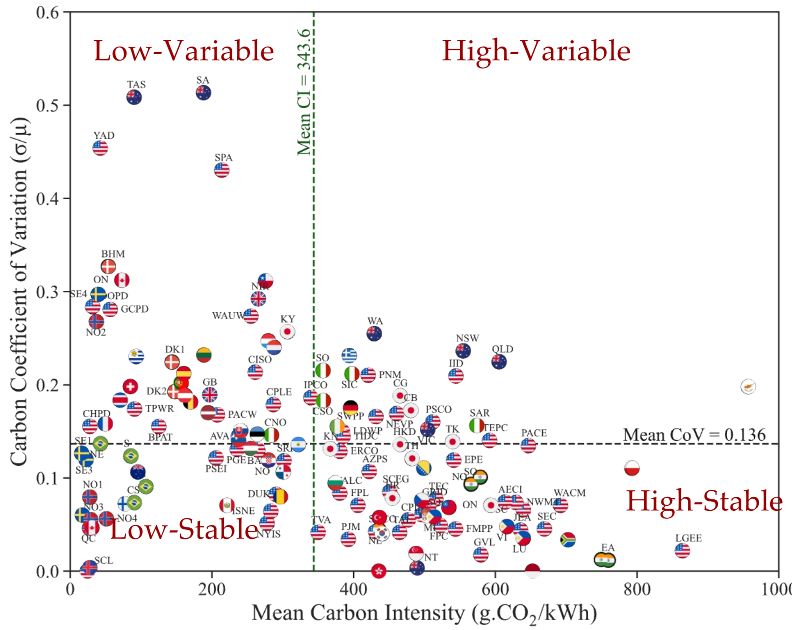


June 15th to 20th of 2022.

Energy's Carbon Intensity (g.CO_2/kWh)

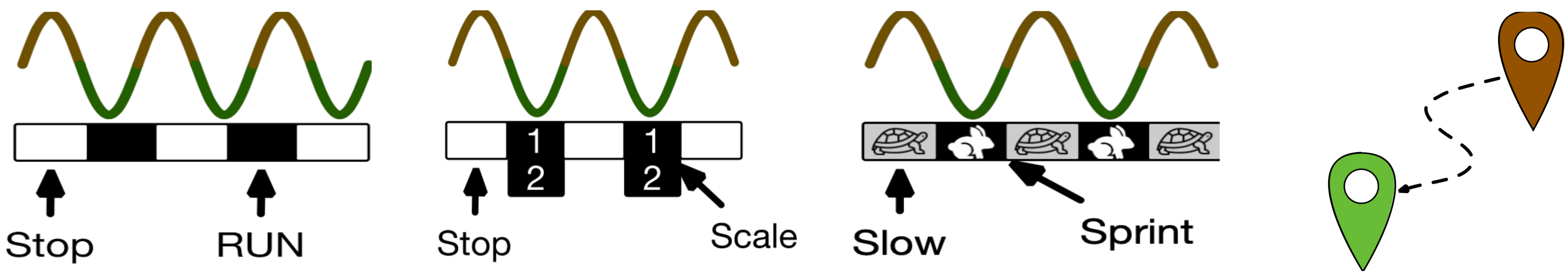


- Carbon Intensity varies temporally by 2×.
- Carbon Intensity differs by ~600 g.CO_2/kWh.
- Electricity grids can be ranked using carbon intensity and variation.



Carbon-aware Resource Management

Computing is equipped with flexibility mechanisms.

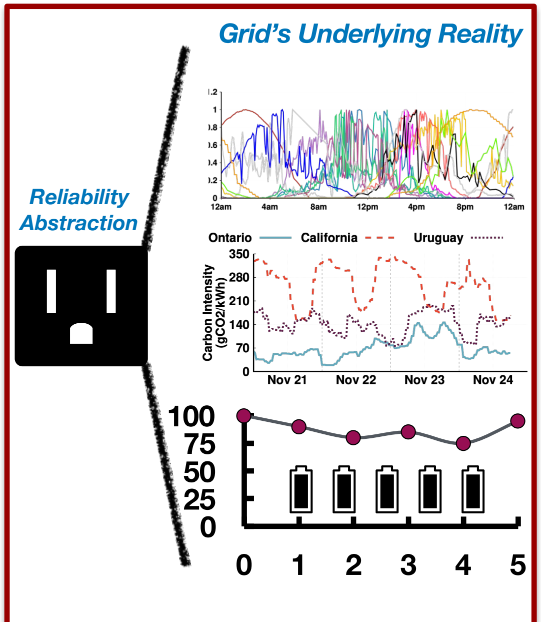


Temporal Shifting | Resource Scaling | Rate Scaling | Spatial Shifting

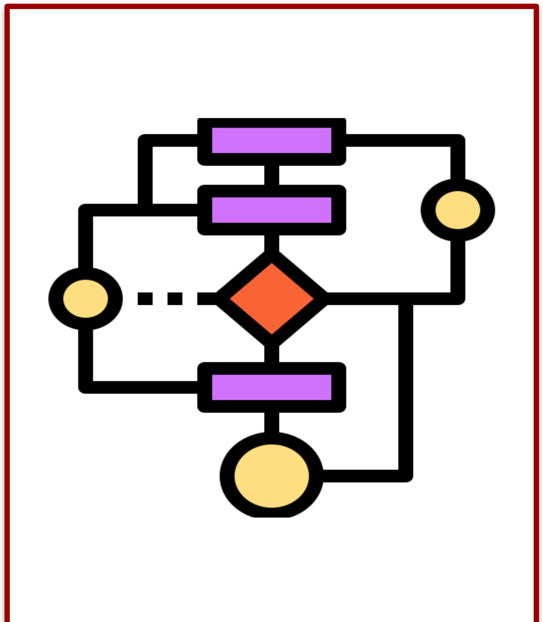
Carbon-aware Computing

How to leverage the carbon intensity variation and computing flexibility?

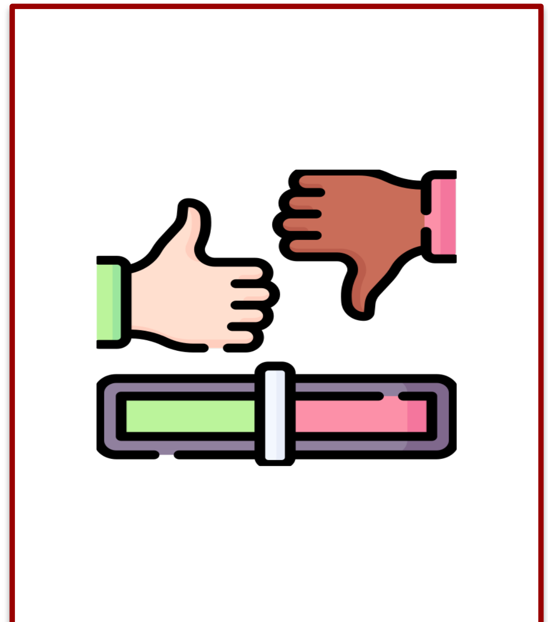
Implementing Carbon-aware Resource Management



Visibility



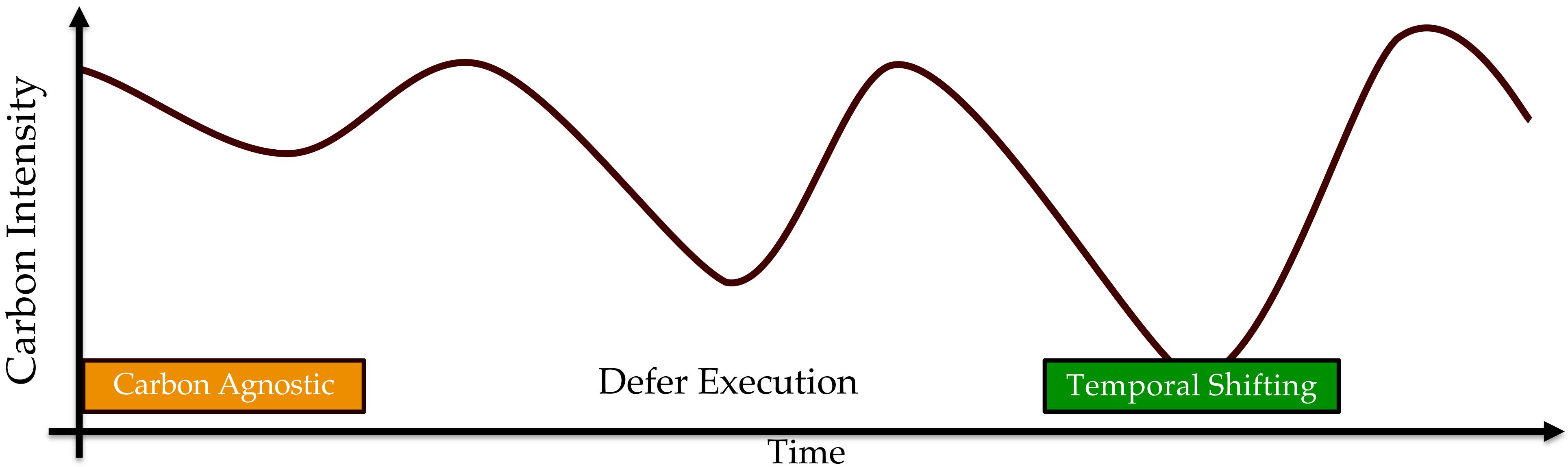
Control



Incentives

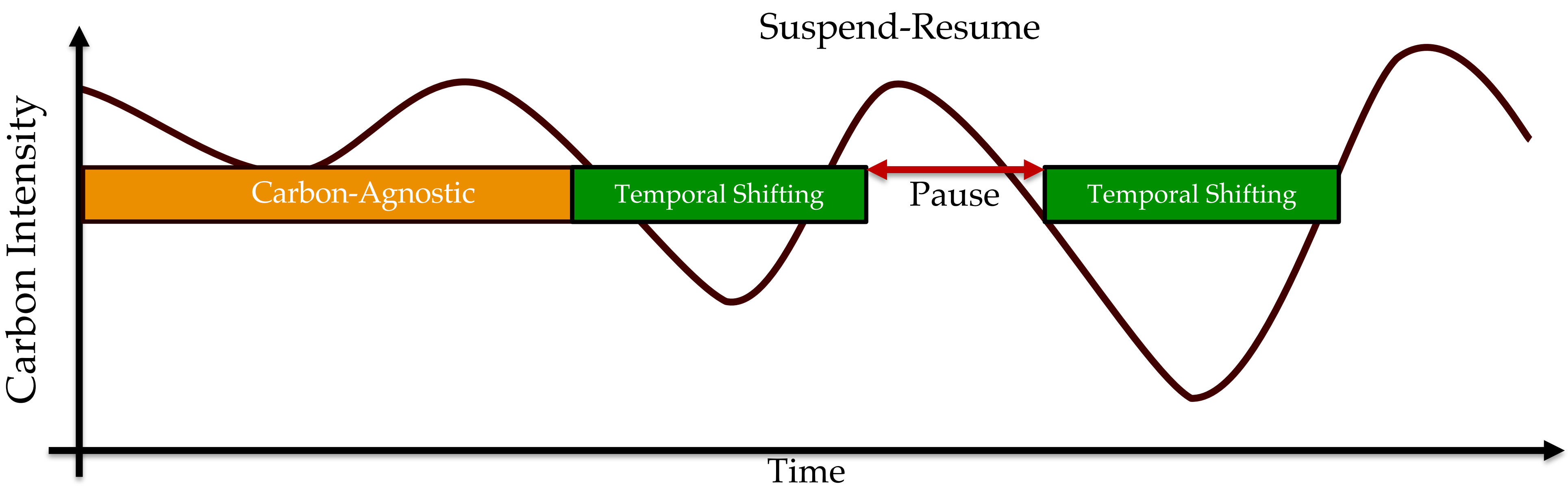
Carbon-aware Temporal Shifting

- Match the availability of low-carbon energy with computing demand.



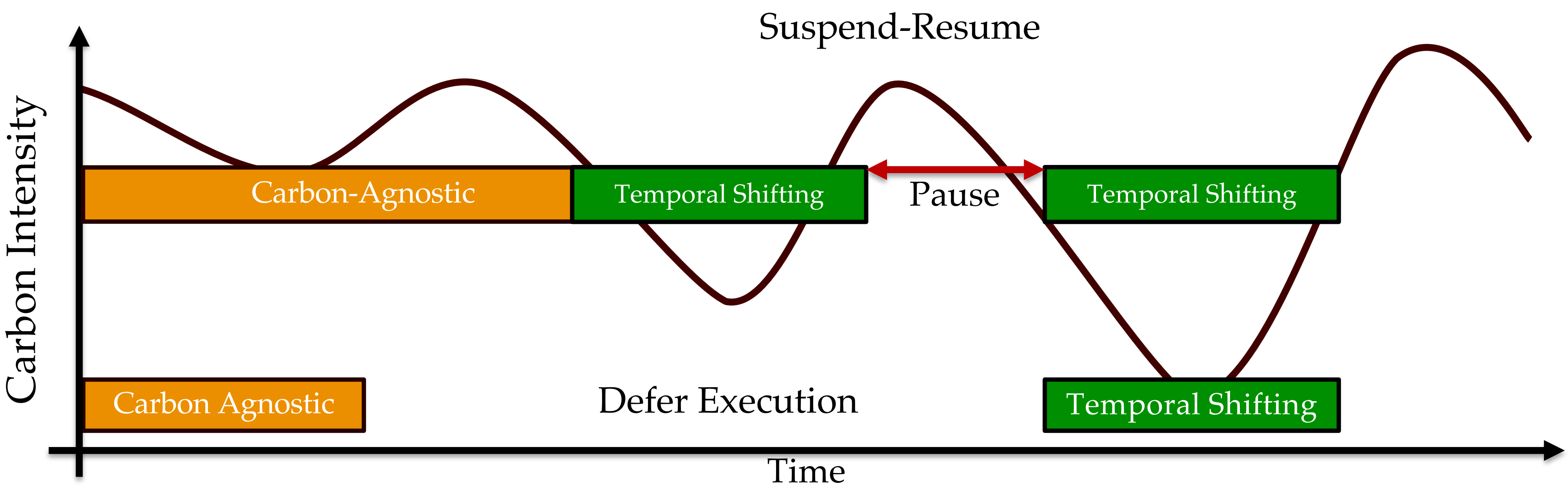
Carbon-aware Temporal Shifting

- Match the availability of low-carbon energy with computing demand.



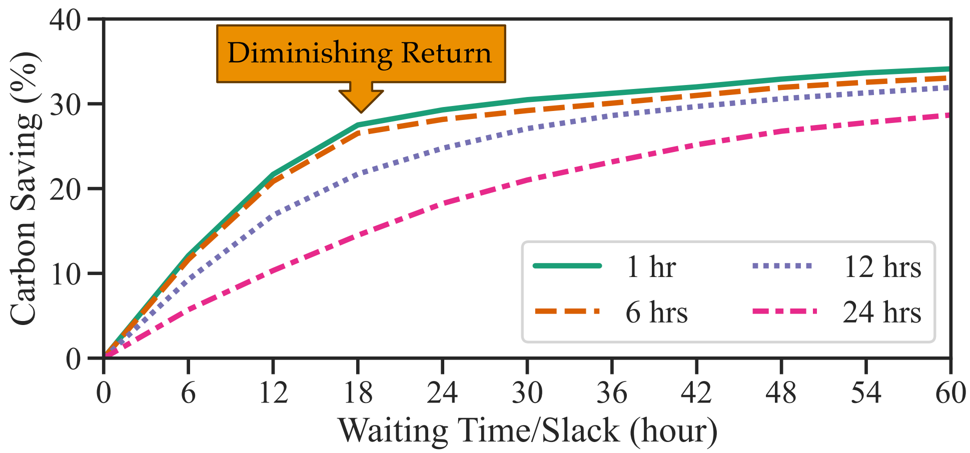
Carbon-aware Temporal Shifting

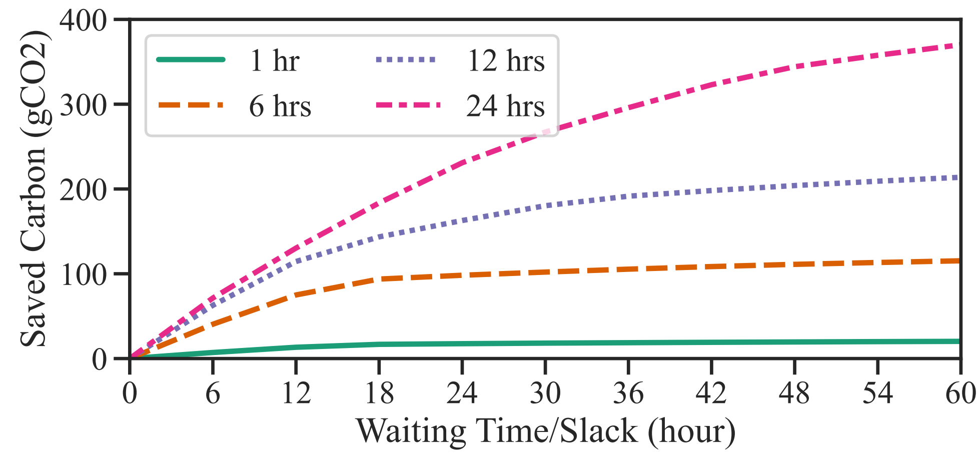
- Match the availability of low-carbon energy with computing demand.



Carbon-aware Temporal Shifting

- Batch Workload: 1hr – 24hrs
- California Carbon Intensity
- Power Consumption (0.2 kWh)
- Suspend Resume (Let's Wait Awhile, Wiesner et al.)

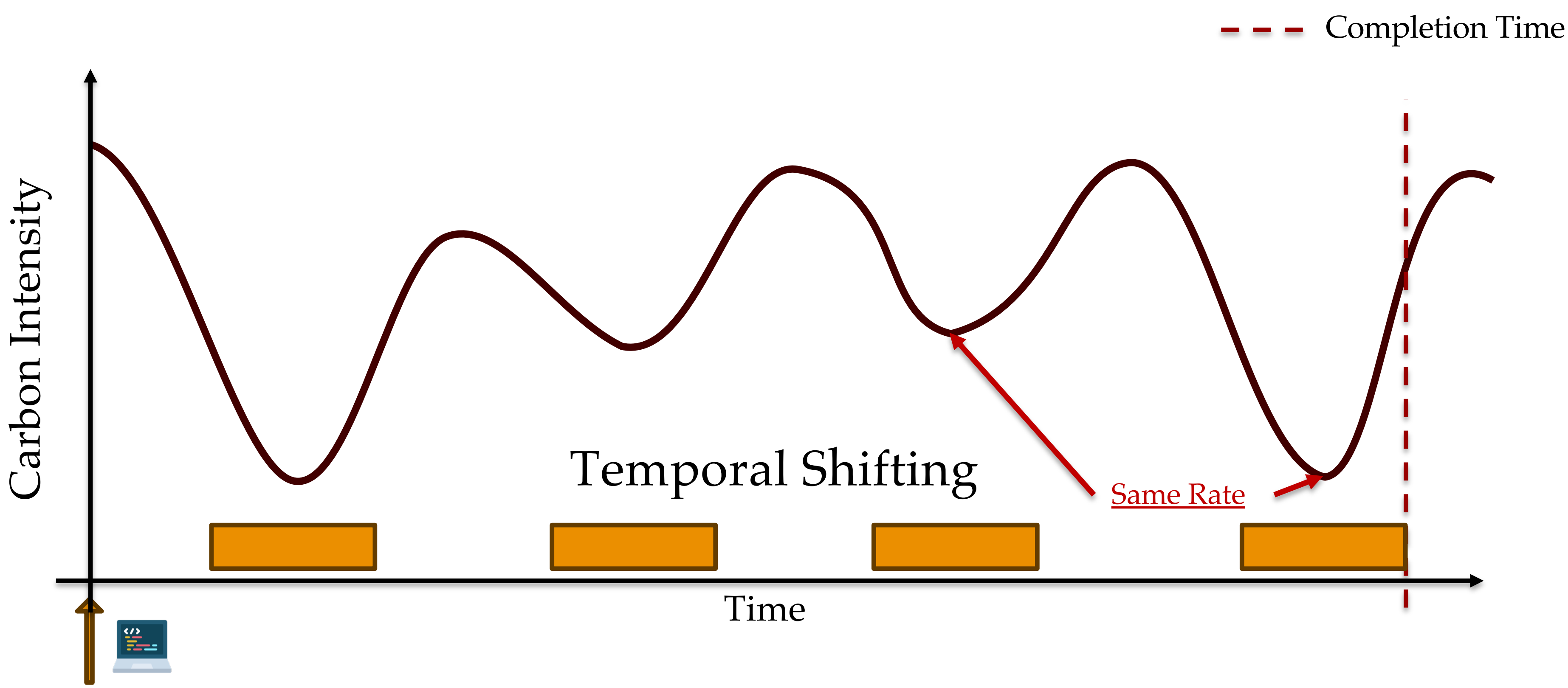




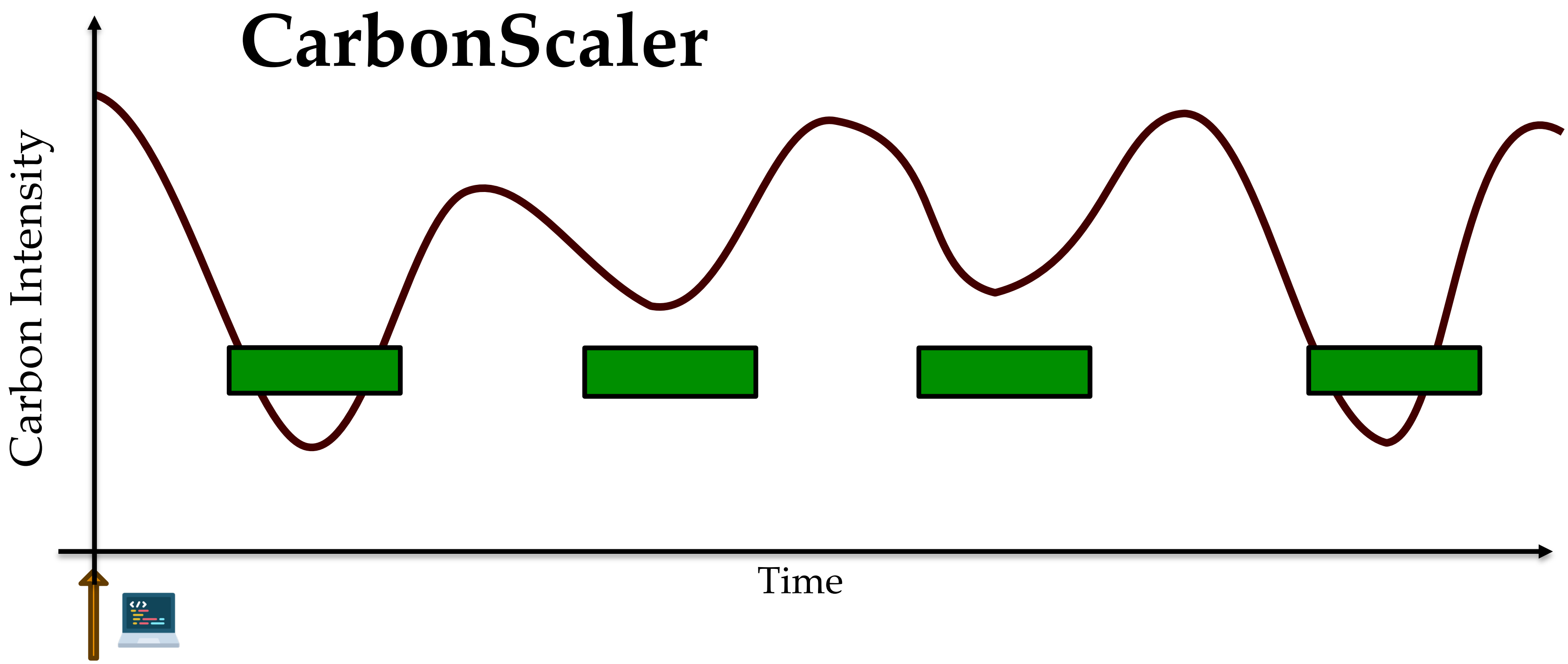
Temporal shifting depends on flexibility in completion time but introduces diminishing returns.
Longer jobs have lower relative savings but higher absolute savings.

UMassAmherst | Manning College of Information & Computer Sciences

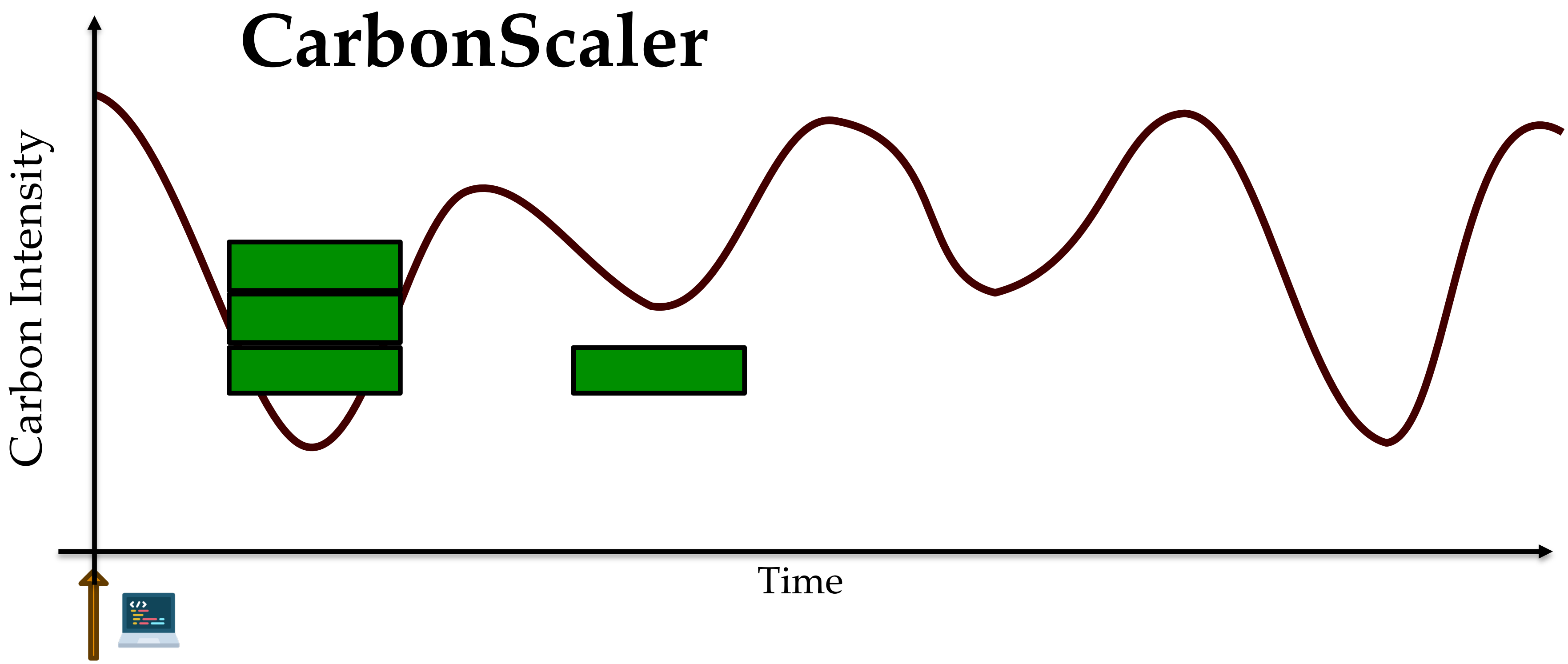
Temporal Shifting and Resource Scaling



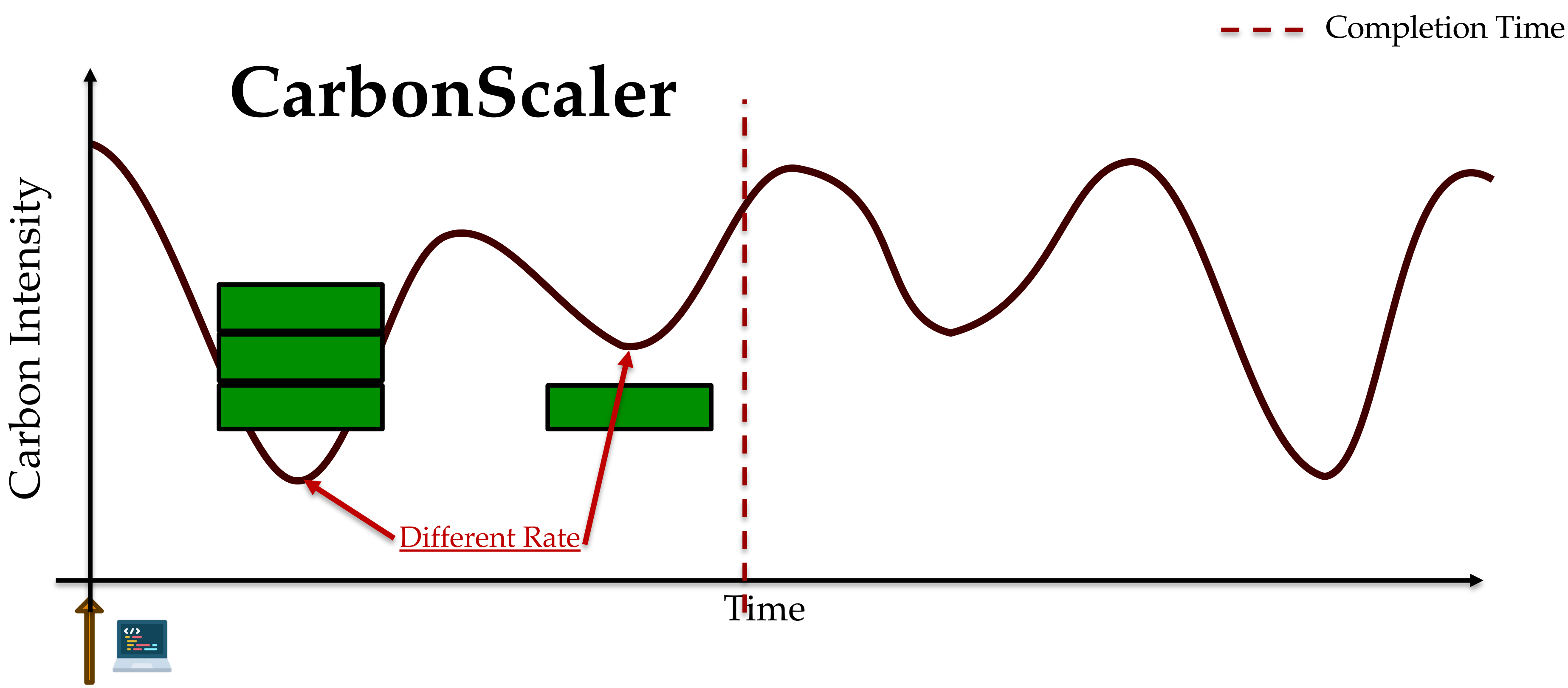
Temporal Shifting and Resource Scaling



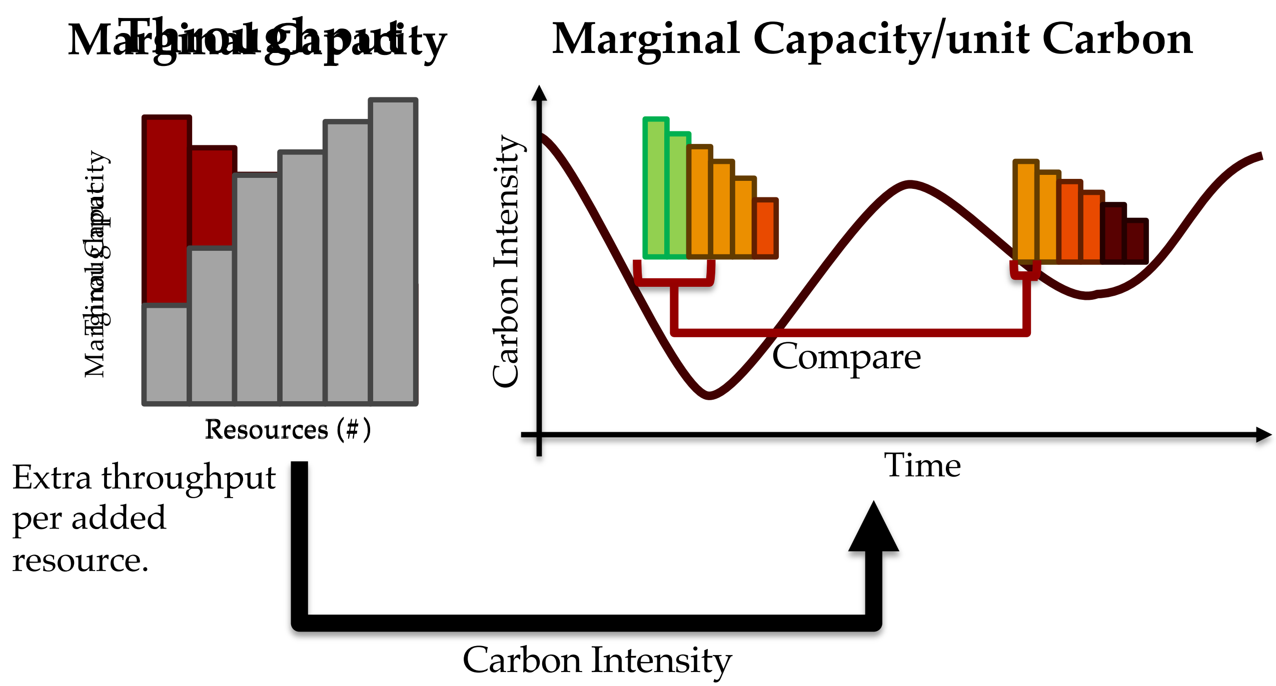
Temporal Shifting and Resource Scaling



Temporal Shifting and Resource Scaling



CarbonScaler Algorithm



Carbon Scaling Problem

Marginal Resource allocation, where greedy algorithms are optimal [1].

Algorithm 1: `Carbon Scaling Algorithm()`

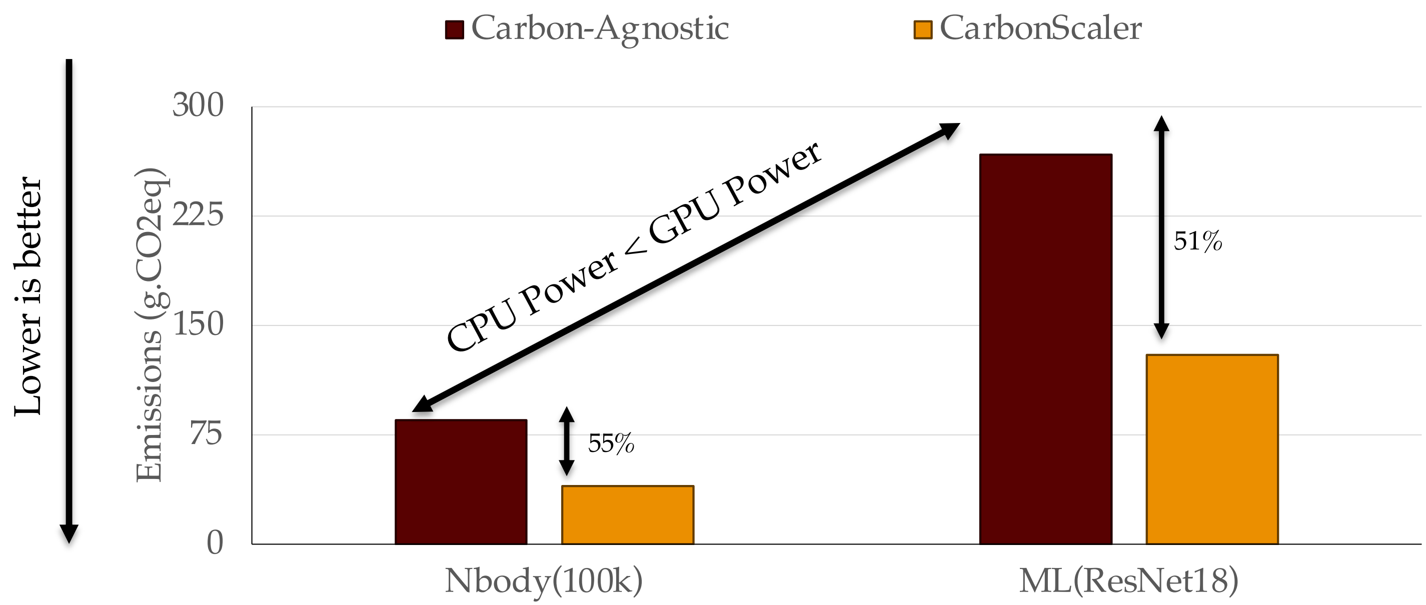
Input: Marginal capacity (MC), time slots $[t, T]$, carbon cost forecast (c), total work (W)
Output: Execution Schedule S

$S \leftarrow [0..0]$;
$L \leftarrow []$;
for $i \in [t, T]$ **do**
  for $j \in [m, M]$ **do**
    L.append($i, j, MC_j/c_i$);
$L \leftarrow$ Sort(L) ; `// w.r.t. Norm. Marginal Cap.`
$w \leftarrow 0$;
while $w < W$ **do**
  $i, j, * \leftarrow L.pop()$; `// next highest` MC_j/c_i
  $S[i] = j$; `// increase allocation in slot` i
  $w.update(S)$;
return S

[1] Federgruen et. al.1986. The Greedy Procedure for Resource Allocation Problems: Necessary and Sufficient Conditions for Optimality. Operational Research.

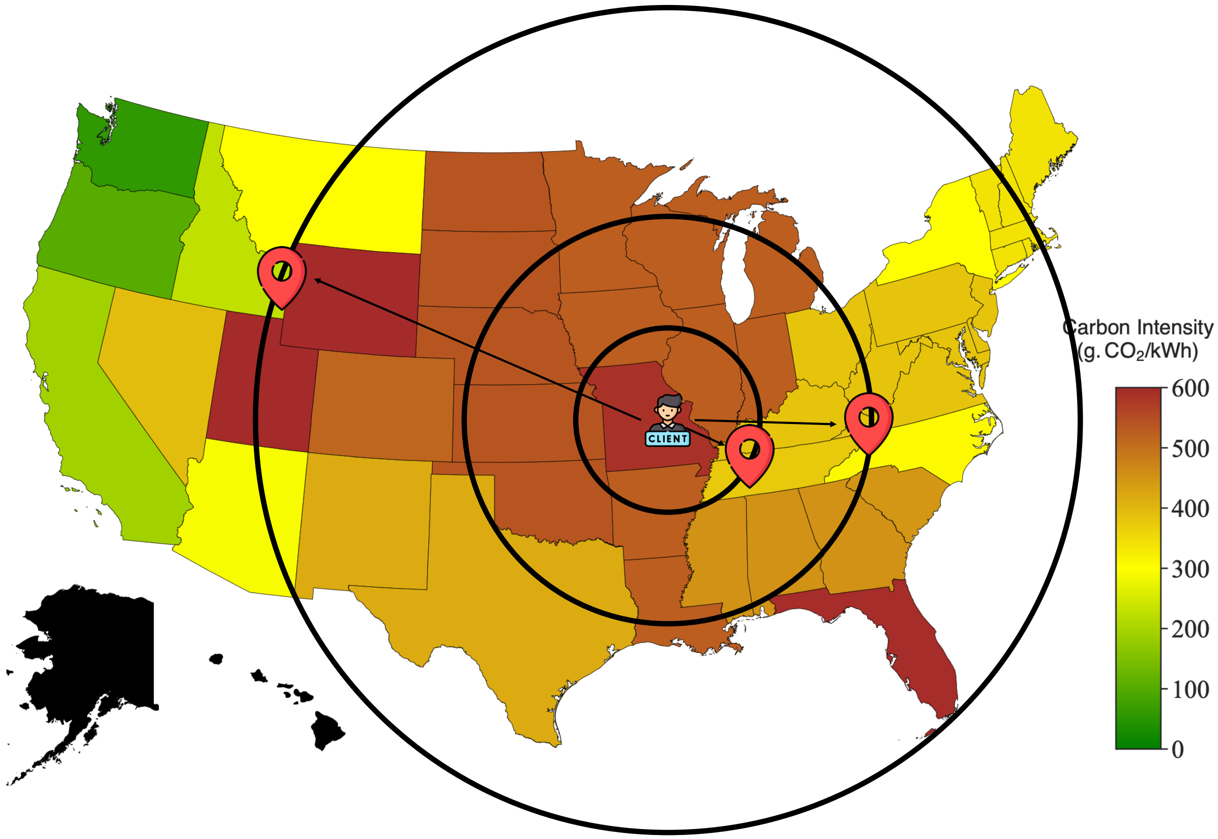
Impact of Workload Elasticity

- Ontario carbon intensity.
- 24hr jobs – No slack.



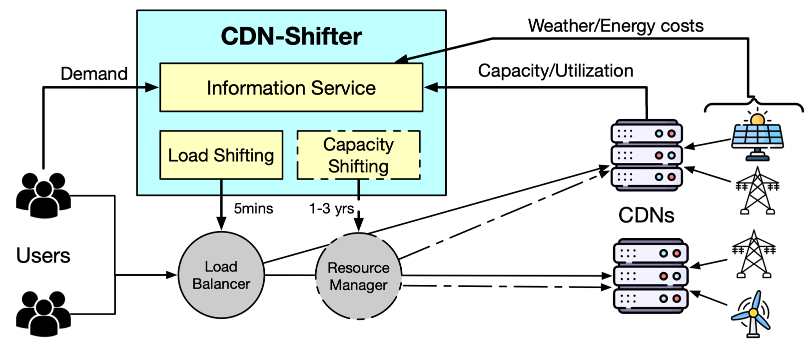
CarbonScaler reduces carbon emissions without increasing completion time.

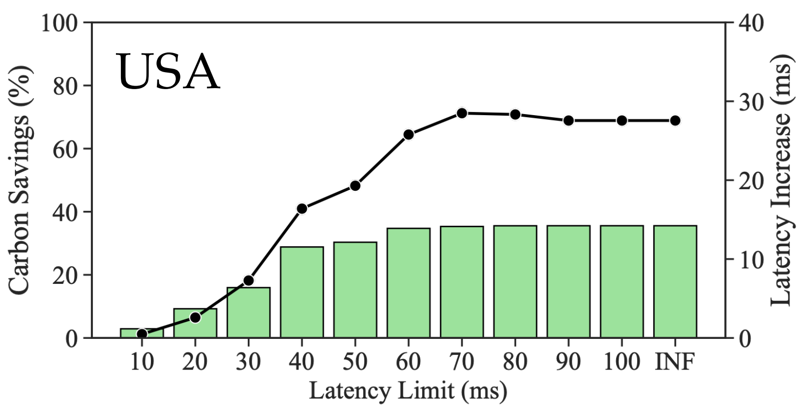
Spatial Shifting

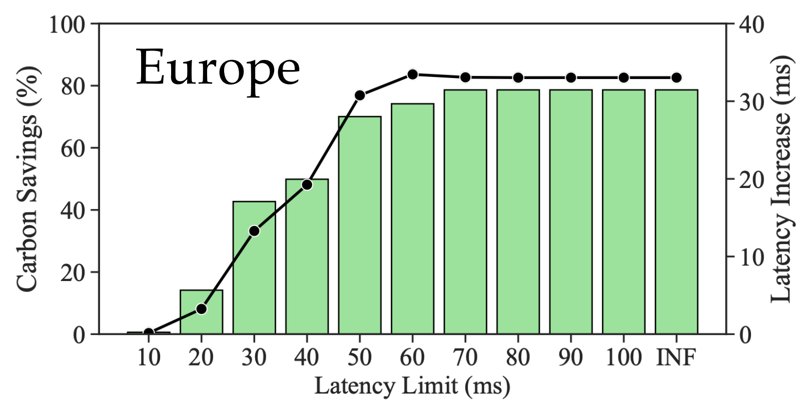


CDN-Shifter (Murillo et al.)

- A carbon- and cost-aware load-shifting framework.
- Capacity Shifting Approach.
- Integrating Solar Energy

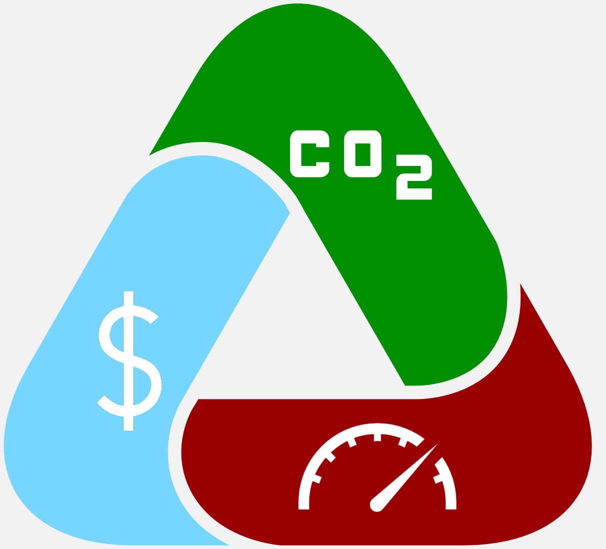






Higher latency leads to higher savings but with diminishing returns.

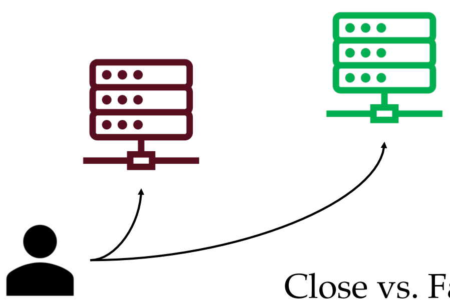
No Free Lunch… Only Trade-offs!



Trade-offs of Carbon-aware resource management

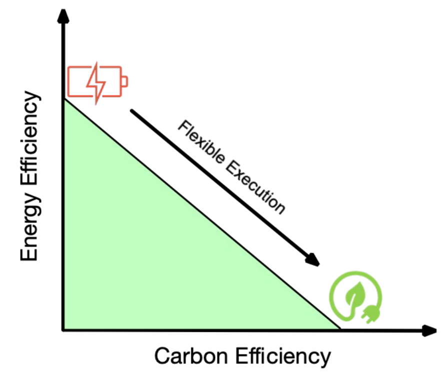
Carbon-aware resource management brings many trade-offs.

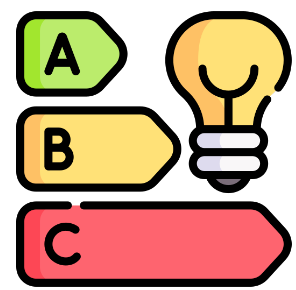
Performance



Close vs. Far
Now vs. Later

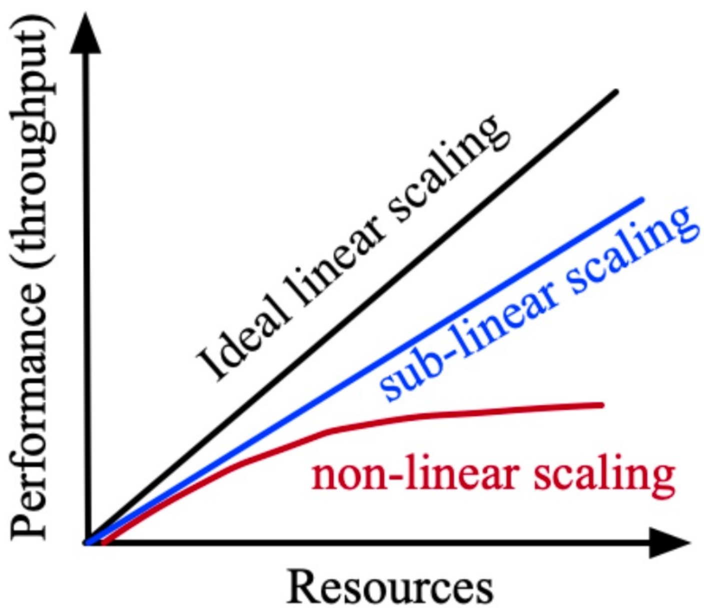
Energy





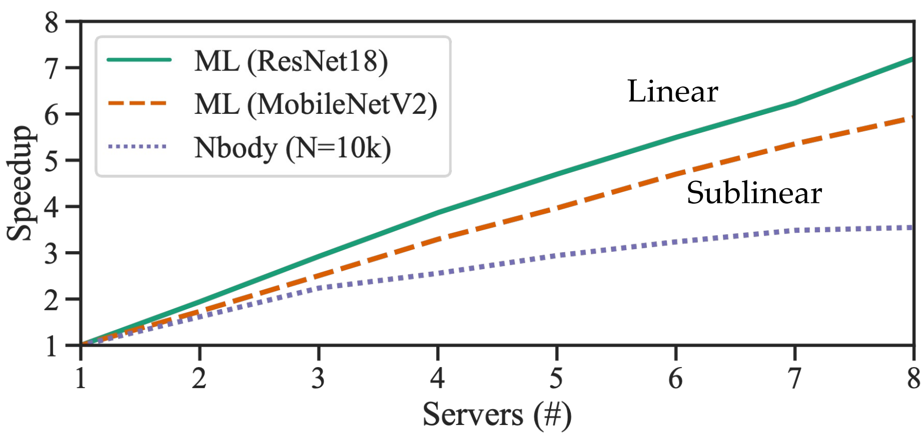
Carbon-Energy Trade-offs

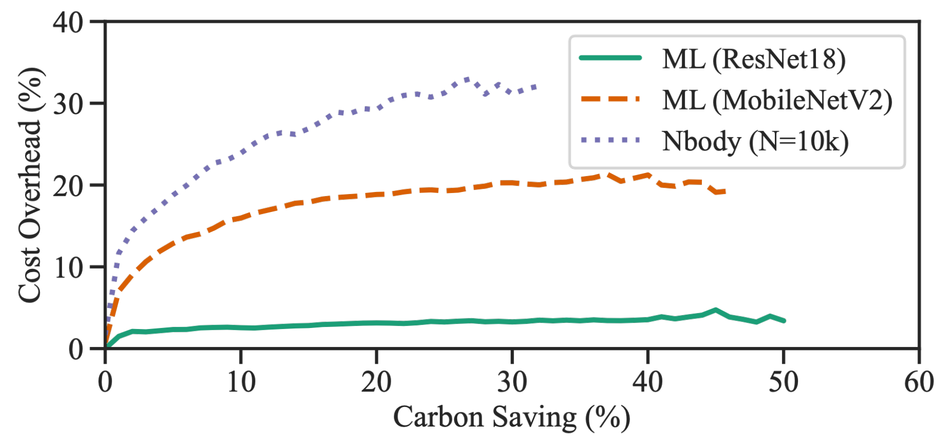
- Elastic Scaling is not free.
- Running at a higher scale is less energy-efficient.



Carbon-Energy Trade-offs

- Elastic Scaling is not free.
- Running at a higher scale is less energy-efficient.

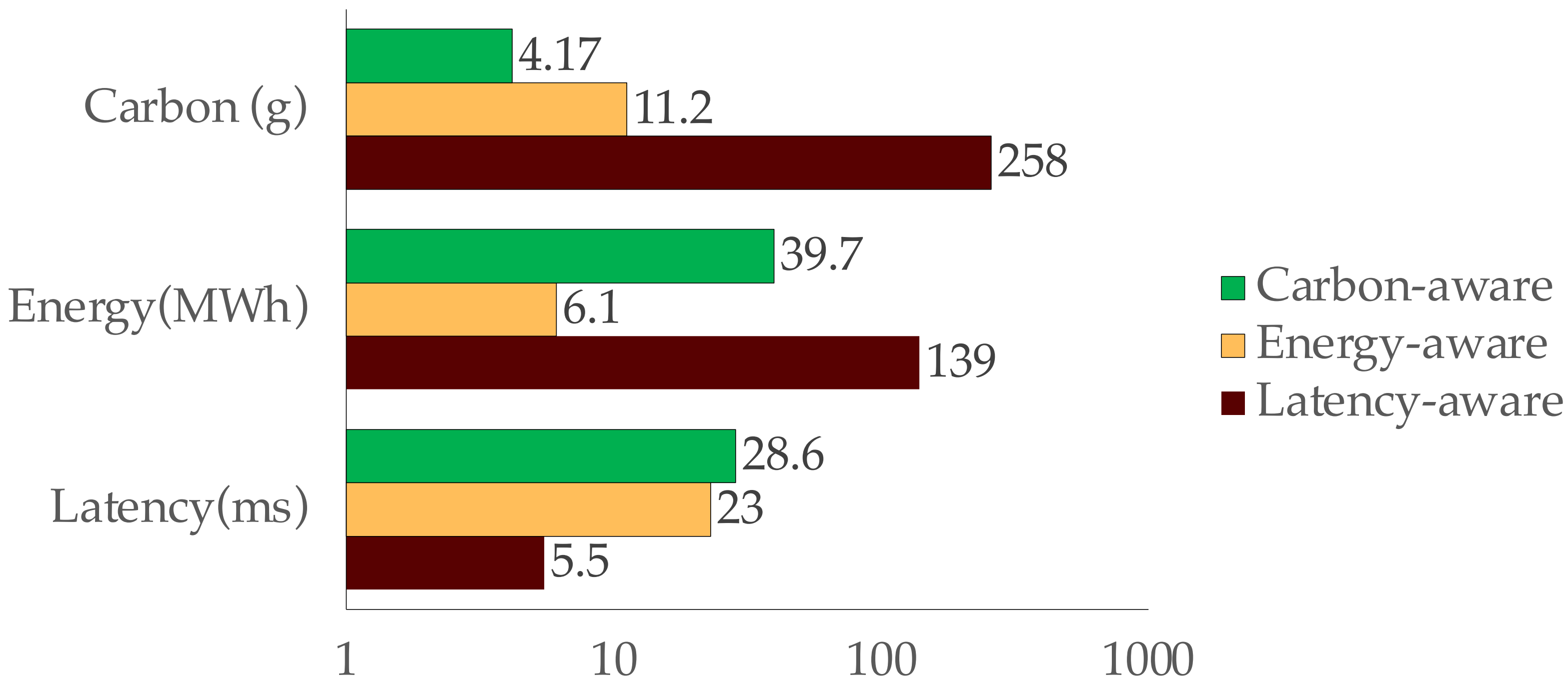




Scaling increases carbon efficiency, and the decrease in energy efficiency depends on workload's elasticity.

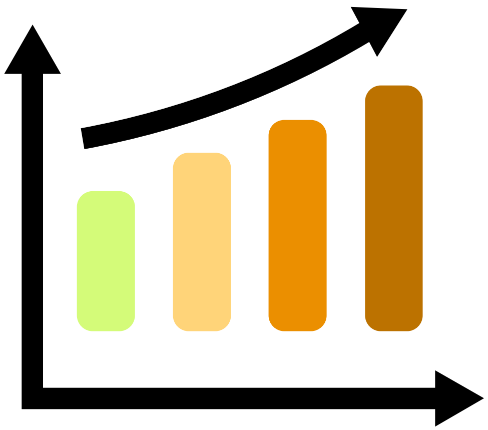
Performance-Energy-Carbon Trade-offs

- DNN placement on heterogeneous edge resources[1].



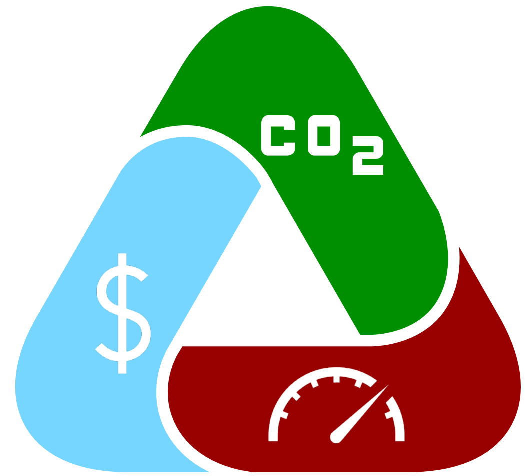
[1]Wu et. al. CarbonEdge

Summary



Carbon footprint of Computing is rising

Lifecycle Emissions



No Free Lunch

Thanks

Walid A. Hanafy
PhD Candidate - UMass Amherst
whanafy@cs.umass.edu

UMassAmherst | Manning College of Information & Computer Sciences

COMPUTING FOR THE COMMON GOOD

Walid A. Hanafy
PhD Candidate - UMass Amherst
whanafy@cs.umass.edu