Sustainable Computing

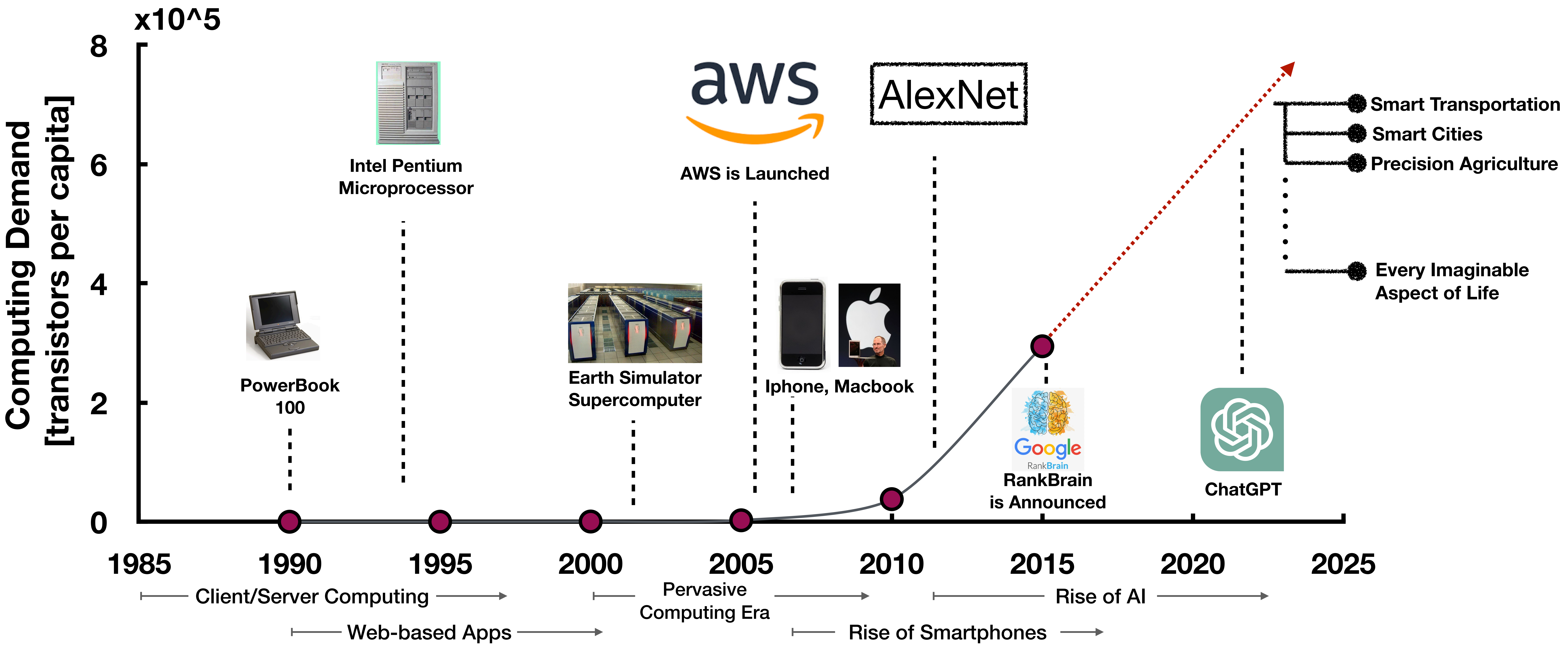
& Computing for Sustainability

Noman Bashir - MIT
nbashir@mit.edu
02/26/2024


First Author
Co-Author
Under-review
Overcommitment [EuroSys'21]
Cost vs. Wait Time [SC'20] [SoCC'21]
Power Monitoring [Performance'23]
Ectothermic Systems [e-Energy'23]
Vision
Sust. Clouds [SoCC'21]
Sust. Computing [HotCarbon'22]
Efficiency War [HotCarbon'23]
Insights
Limits & Benefits [EuroSys'24]
Avg. vs. Marg. [e-Energy'24]
Green Mirage [e-Energy'24]
Algorithms
DPTR, RORO [SIGMETRICS'24 x 2]
DANISH [In prep]
LACS [e-Energy'24]
CLIP [ICML'24]
Systems
GreenSlurm [ASPLOS'24]
EcoVisor [ASPLOS'23]
LXC-C [SoCC'23]
CarbonScaler [SIGMETRICS'24]
Embodied Estimation [In preparation]
MIT Cluster Carbon Footprint [In preparation]
Embodied Accounting [HotCarbon'23]
Decarbonize Buildings [KDD'24]
Decarbonize Buildings [COMPASS'20] [e-Energy'19]
Decarbonize Buildings [BuildSys'22] [e-Energy'23]
Solar Integration in Grid [BuildSys'17,18 20]
Green Ridesharing [e-Energy'24]
Gen-AI Impact Evaluation [MIT Press]
Supply Chain
Extract, Process, Manufacture, Transport
software
algorithms, scheduling
energy
idle, dynamic
carbon
grid, solar, battery
Datacenter/Edge/Device
EOL
End of Life
Apps
embodied
operational
Electric Grid
Buildings
Transportation
Sustainability of Computing
Computing for Sustainability

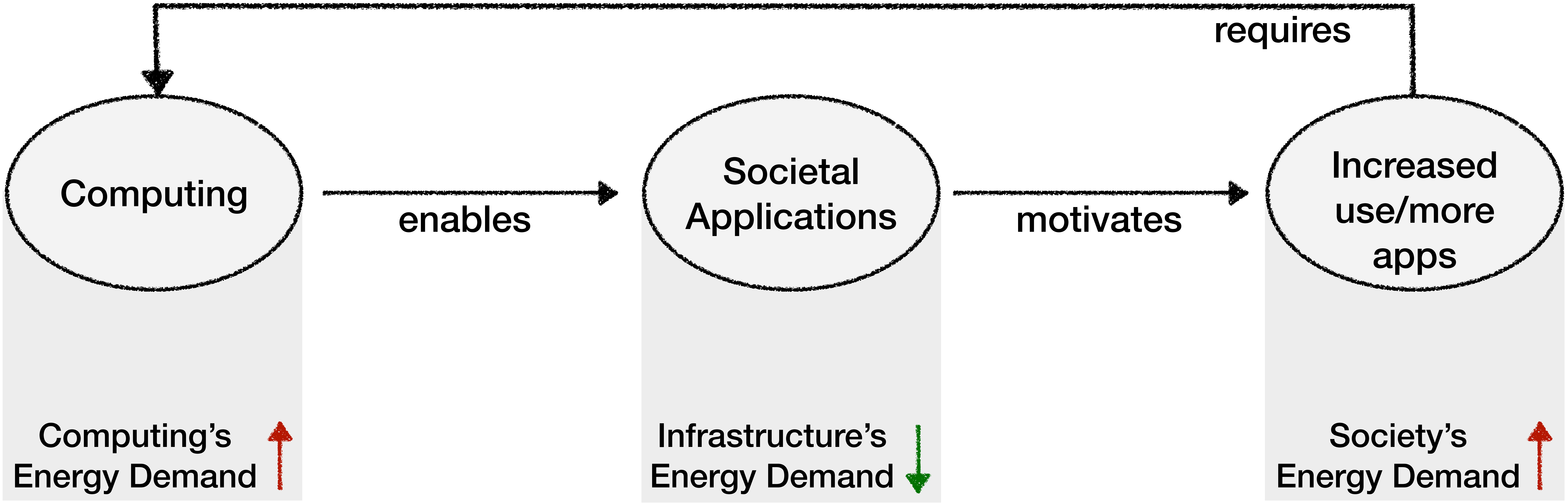
Computing's Demand is Growing Exponentially

- Society continues to find useful applications

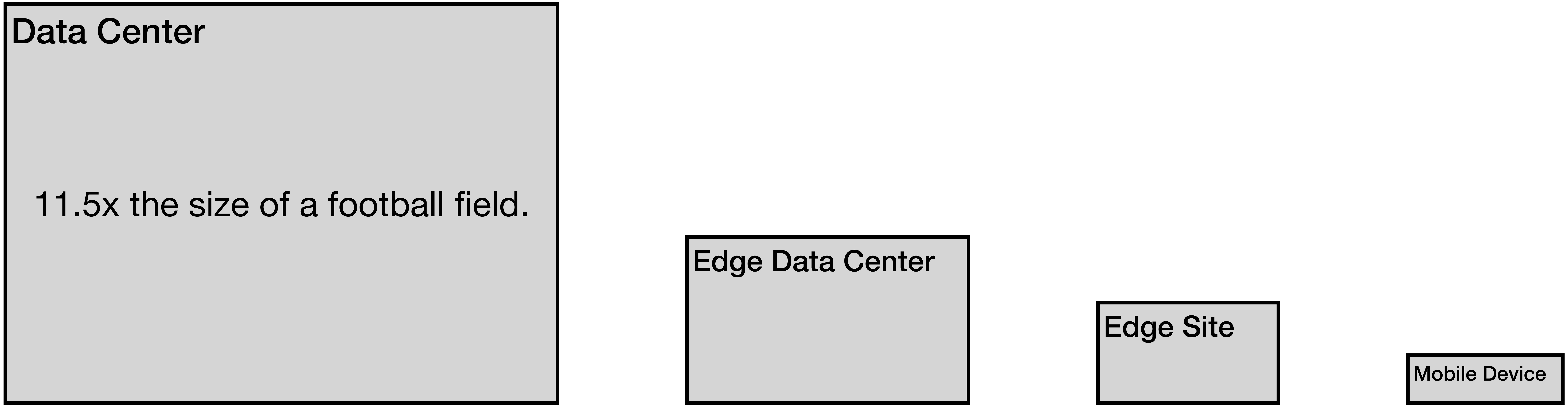


Source: "Unimaginable Output: Global Production of Transistors" - Darrin Qualman

Implications of Increasing Computing Demand



How is Computing Demand Served?



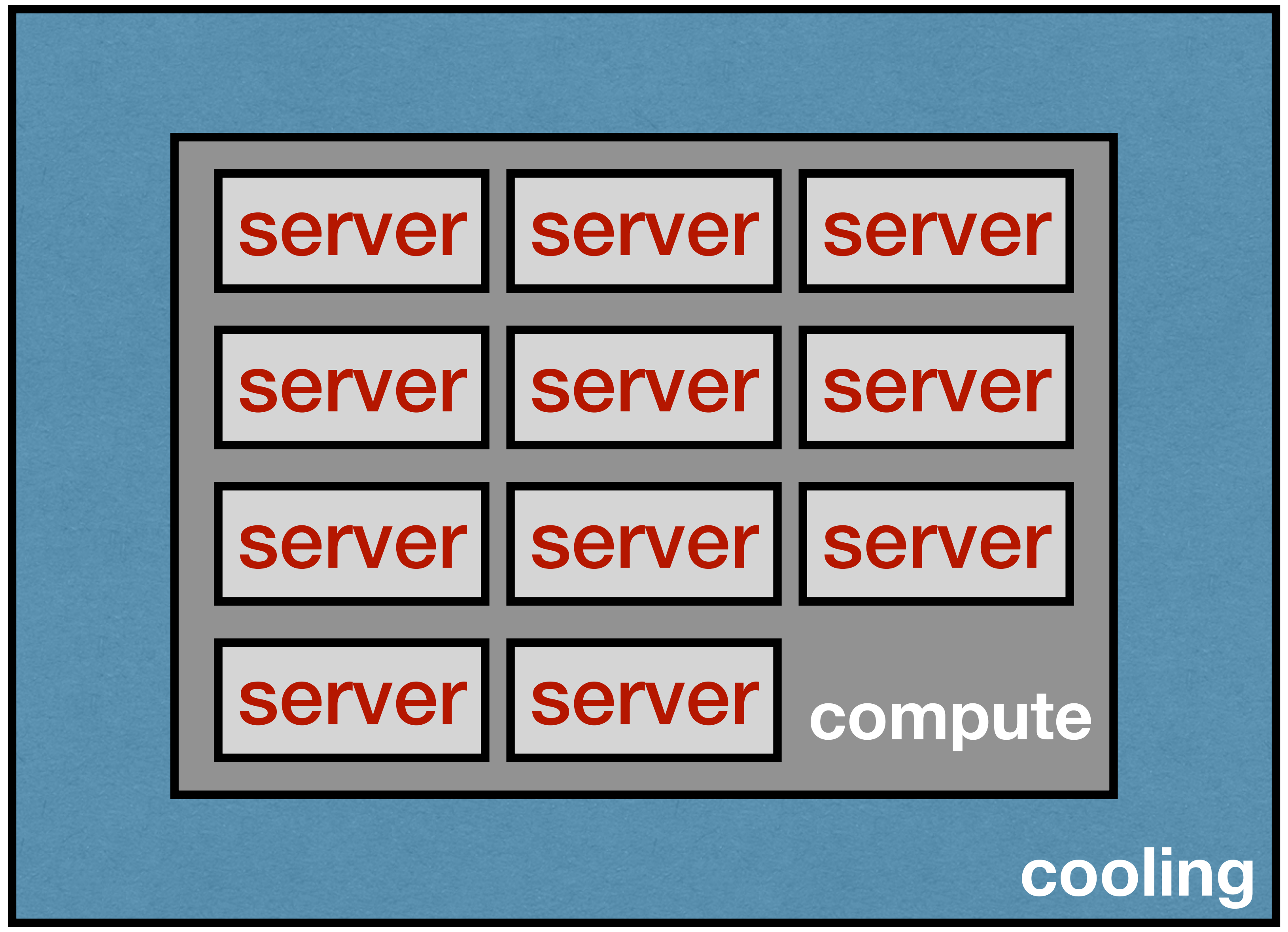
Thousands of servers and data storage, e.g., Google Dalles data center houses **~100k servers** and consumes **100MW** of power (enough for a small city)

10s-100s of servers and data storage,
1,000 sqft to 50,000 sqft
a **few kW** to a **few MW**

Mobile devices and small storage hand-held etc.
a few watts

note: figures are not drawn to scale.

What Contributes to Data Center's Cost, Energy, Carbon Footprint?



Cost

- **Servers:** Cost a lot and are replaced every 3-5 years.
- **Building:** Capital investment, depends on location.
- **Energy:** Major cost of datacenter, depends on location.

Energy

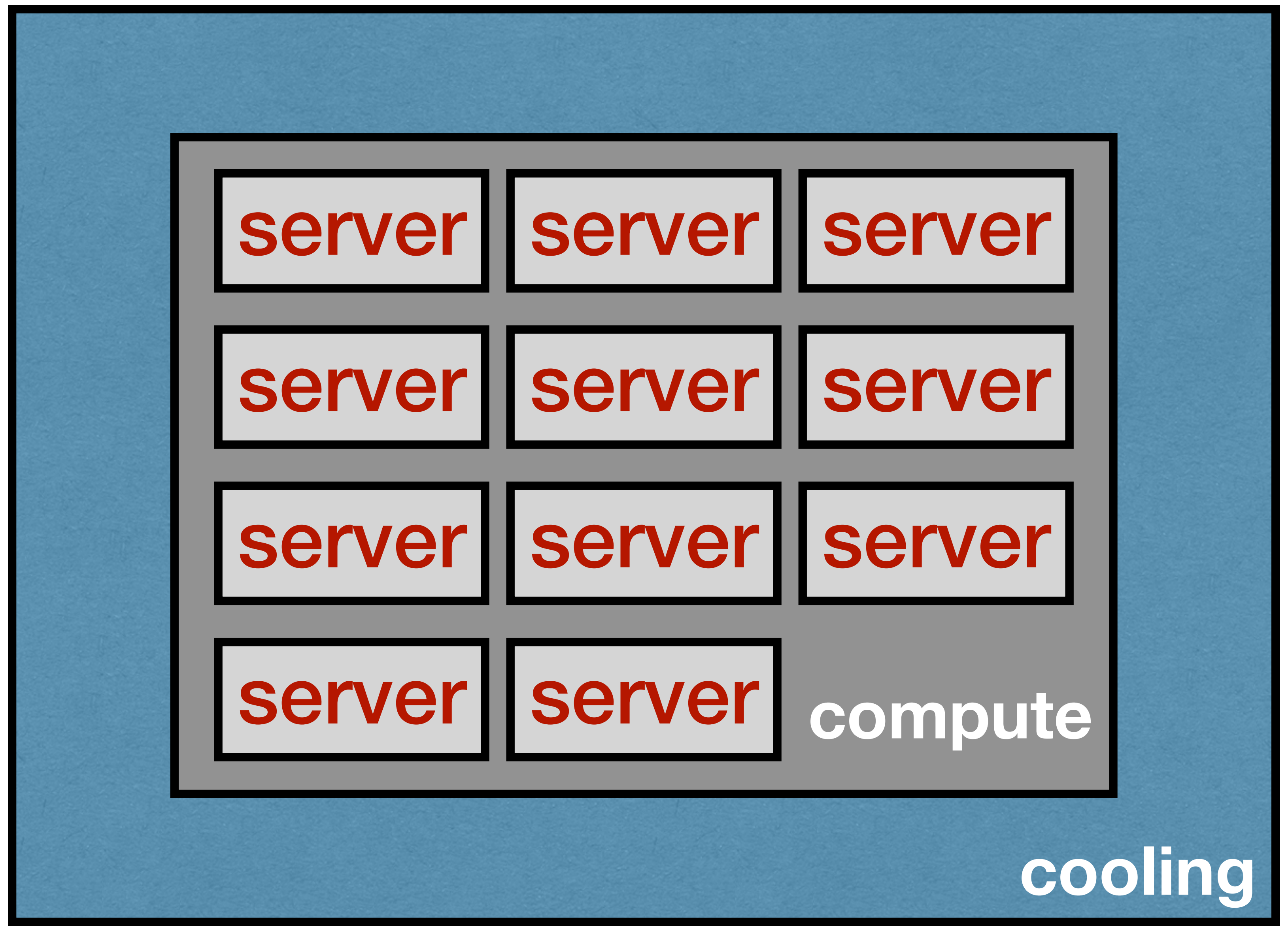
- **Computing:** Become more energy efficient over time.
- **Cooling:** Wasted energy, significantly reduced over years.

Carbon

- **Embodied:** Carbon emissions from manufacturing/building.
- **Operational:** Emissions from energy use for compute and cooling.

How to Serve Computing's Demand in a Sustainable Manner?

Sustainable —> least carbon intensive way.



Carbon

- **Embodied:** Carbon emissions from manufacturing/building.
- **Operational:** Emissions from energy use for compute and cooling.
  - From the energy used to **run** the servers.
  - From the energy used to **cool** the servers.

Reduce Embodied Emissions and Reduce Operational Emissions

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

$$\text{Carbon Footprint} = \frac{\text{10 cycles per inference request} \times \text{100 inference requests}}{\text{5 cycles per kWh} \times \text{1 kWh per gCO2eq}}$$

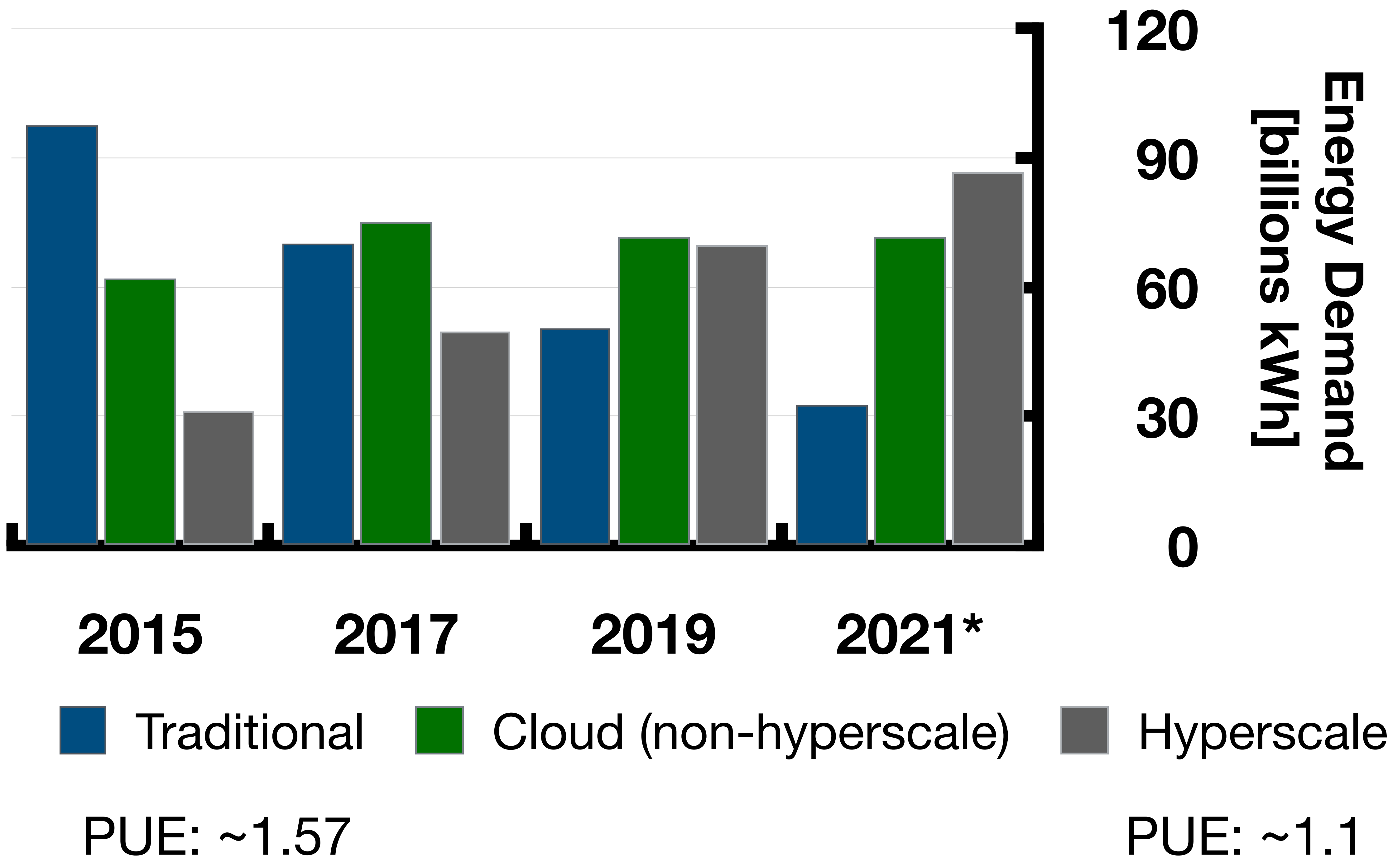
$$\text{Carbon Footprint} = \boxed{\text{200 gCO2eq}}$$

History: Driving Factors Behind Innovations in Data Centers

Cost of Energy Has Been Driving Innovation

- Assume 100,000 servers
- **Monthly cost of 1 server**
  - 500W server
  - Cost = (Watts X Hours / 1000) * cost per kWh
  - Always-on server monthly cost = **$50**
- **Monthly cost of 100k servers = $5M**
- What about the cost of cooling?
  - Use Power Usage Effectiveness (PUE)
  - PUE = 2 —> double the cost
  - **PUE =1.2 —> 10% extra on $5M ($6M)**

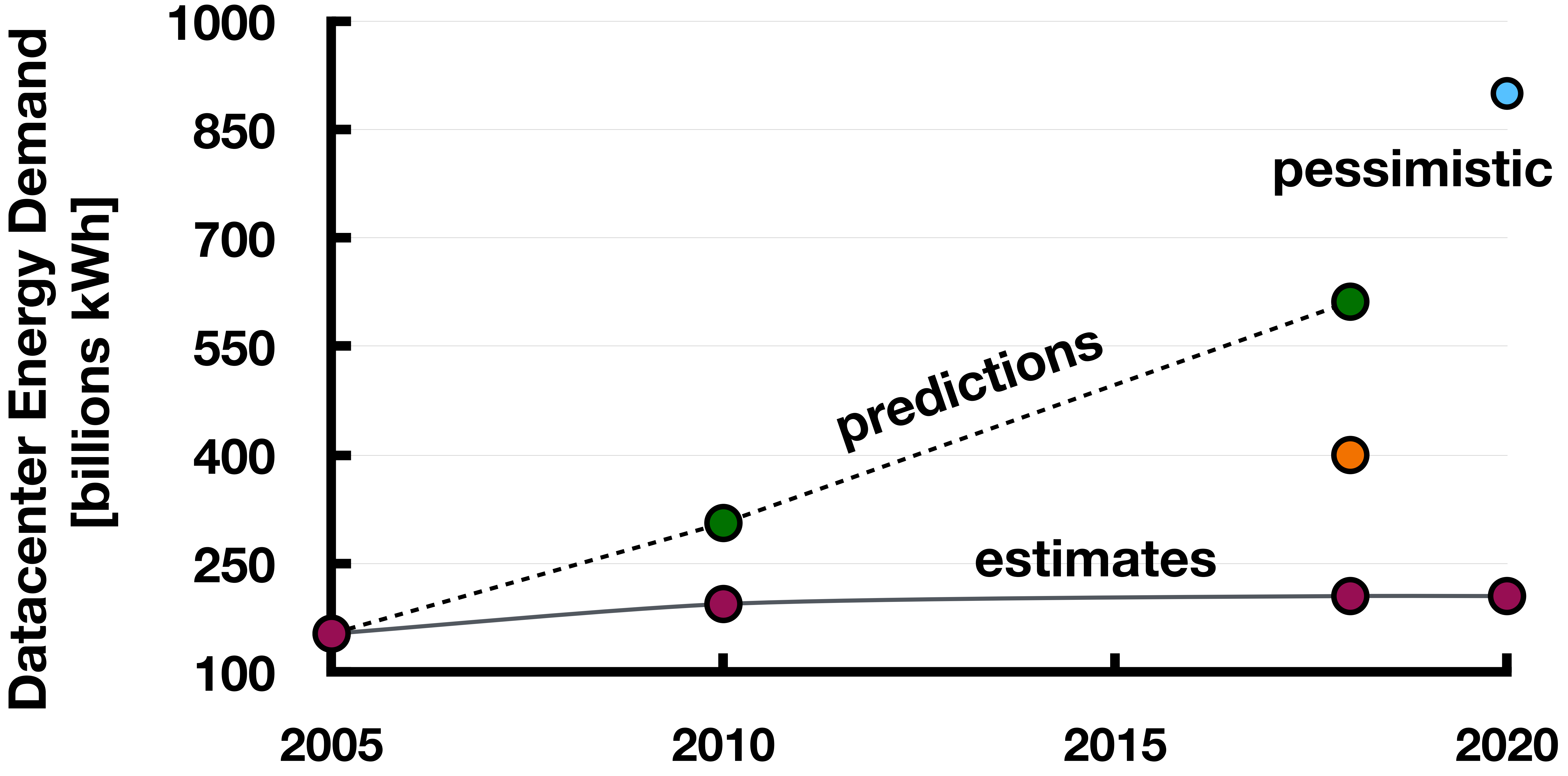
Shift from Traditional Data Centers to Cloud



Source: Global data centre energy demand by data centre type, 2015-2021 - IEA

Energy Efficiency Gains Moving Forward

- Most optimistic estimates suggest 6% increase from 2010-2018



Demand Accelerating
vs
Energy-efficiency Gains
Slowing Down

EPA Report to Congress on Server and Data Center Energy Efficiency (2007)
Recalibrating Global Data Center Energy-use Estimates - Eric Masanet (2020)
Efficiency Gains are Not Enough: Data Center Energy Consumption Continues to Rise Significantly - Ralph Hintemann (2018)

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]

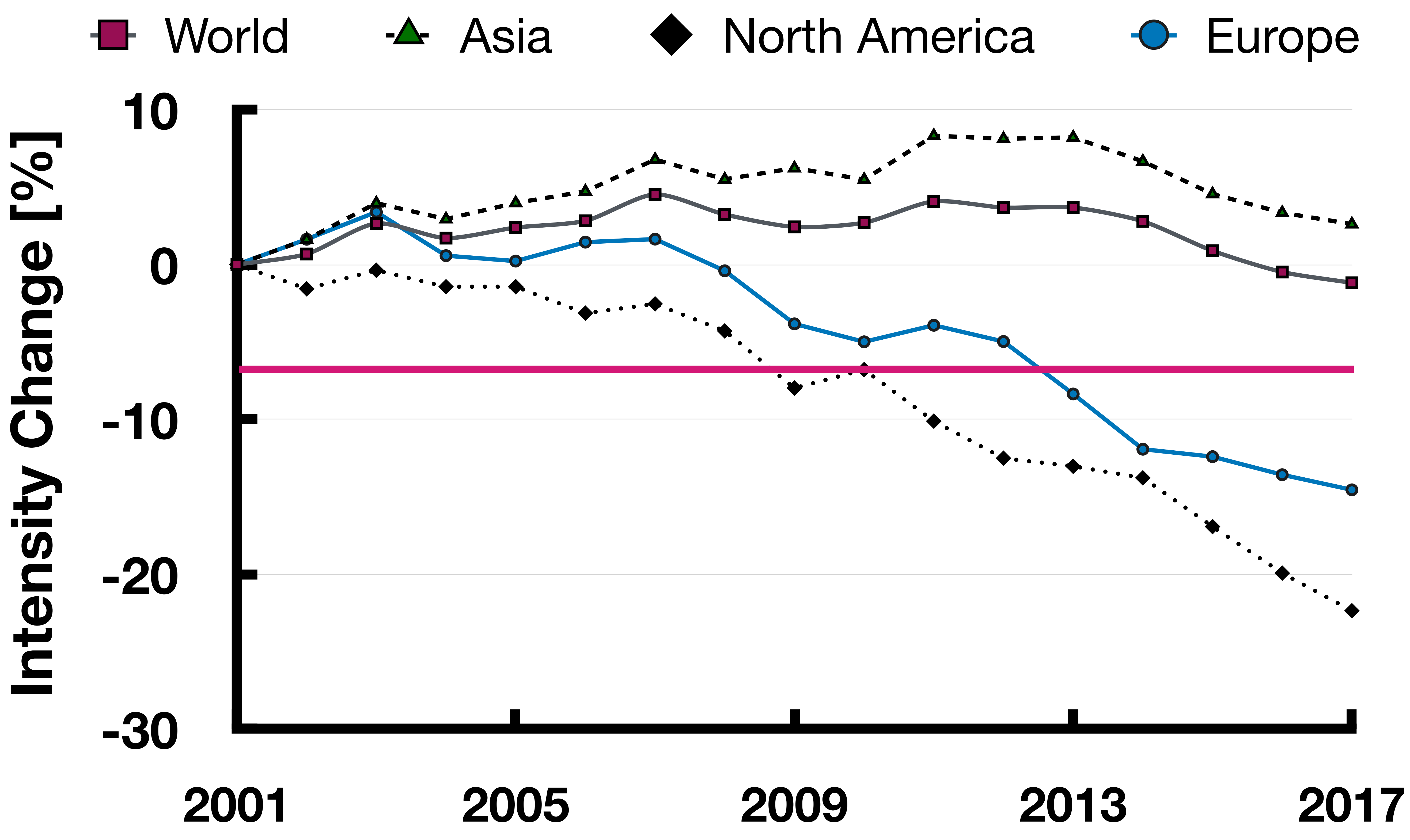
[unbounded]

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

Industry has helped subsidize zero-carbon energy

Grid's Carbon Intensity Has Been Decreasing

- Energy's carbon efficiency in the US has improved by 45.6% over 2001-2017

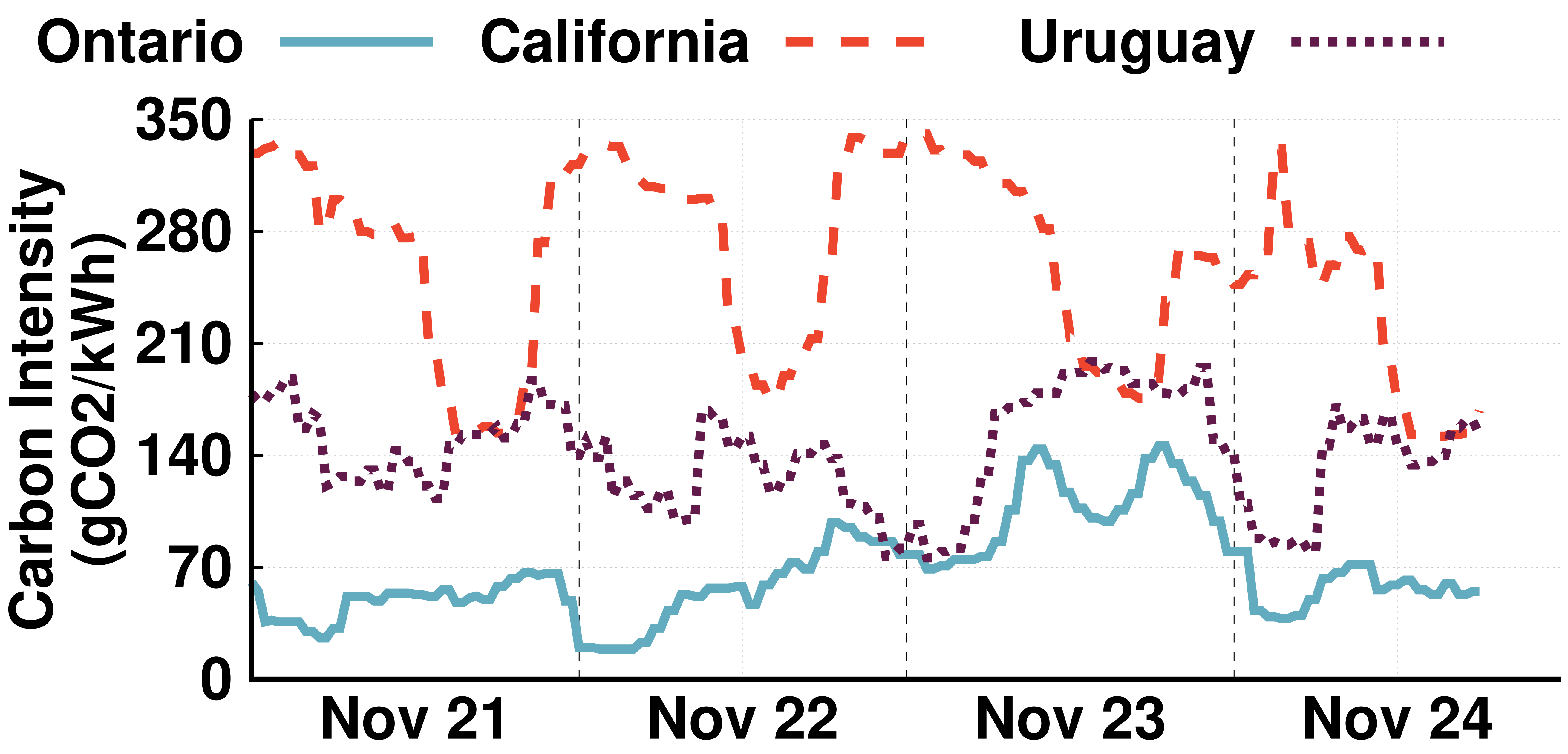


Carbon intensity may never truly reach 0gCO2eq per kWh.

It may actually increase in parts of the world.

Source: Ember Global Electricity Review (2022)
Source: BP Statistical Review of World Energy
Source: Ember European Electricity Review (2022)

Carbon Intensity of Electricity Varies Across Space & Time



Spatial Variations

Move to the greenest data center possible

Temporal Variations

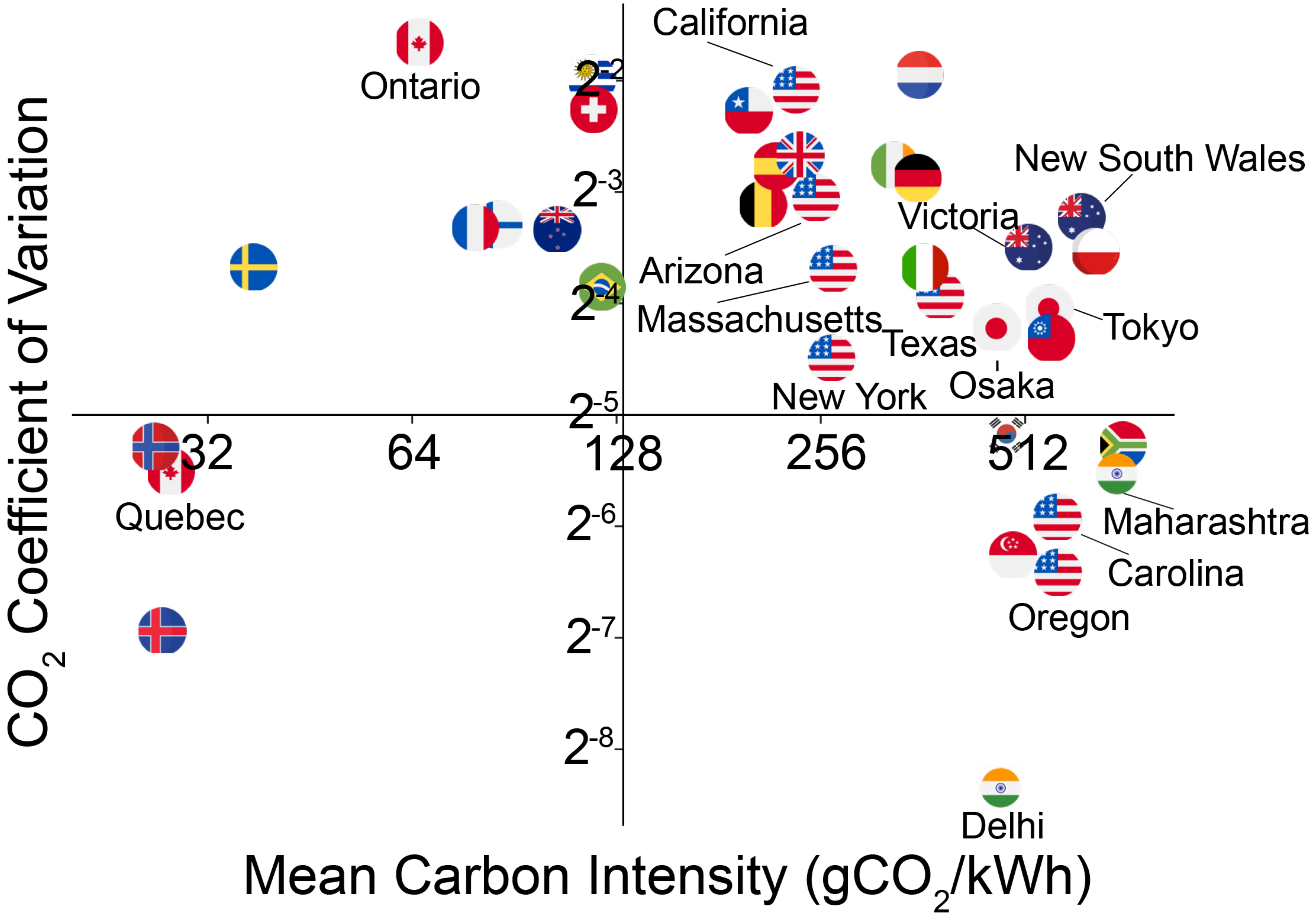
Move to a time slot with the lowest carbon emissions

Run when and where low-carbon energy is available.

Clean Energy is Variable and Unreliable

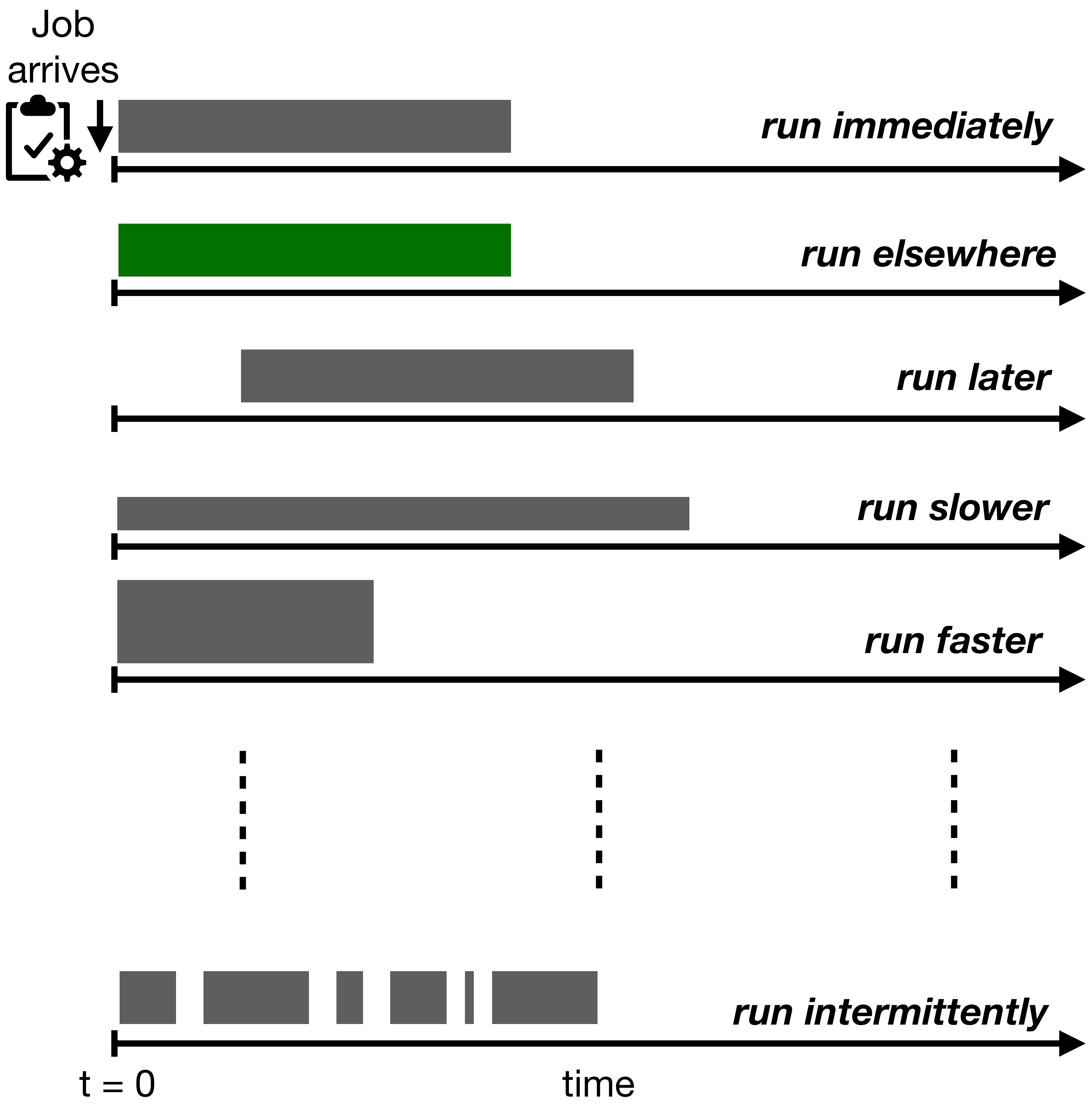
- Carbon intensity variation: **less than 50g** to **more than 800g** across time and geographical regions.

More regions in the world would look like Ontario in near future.



The Good News: Computing's Unique Advantages

Driven by efforts to improve user experience & scale



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

How can we leverage carbon intensity variations and computing's flexibility?

Enabling Sustainable Clouds: The Case for Virtualizing the Energy System

Noman Bashir*, Tian Guo^, Mohammad Hajiesmaili*, David Irwin*, Prashant Shenoy*, Ramesh Sitaraman*, Abel Souza*, Adam Wierman^^

Work published at:

SoCC’21, ASPLOS’23

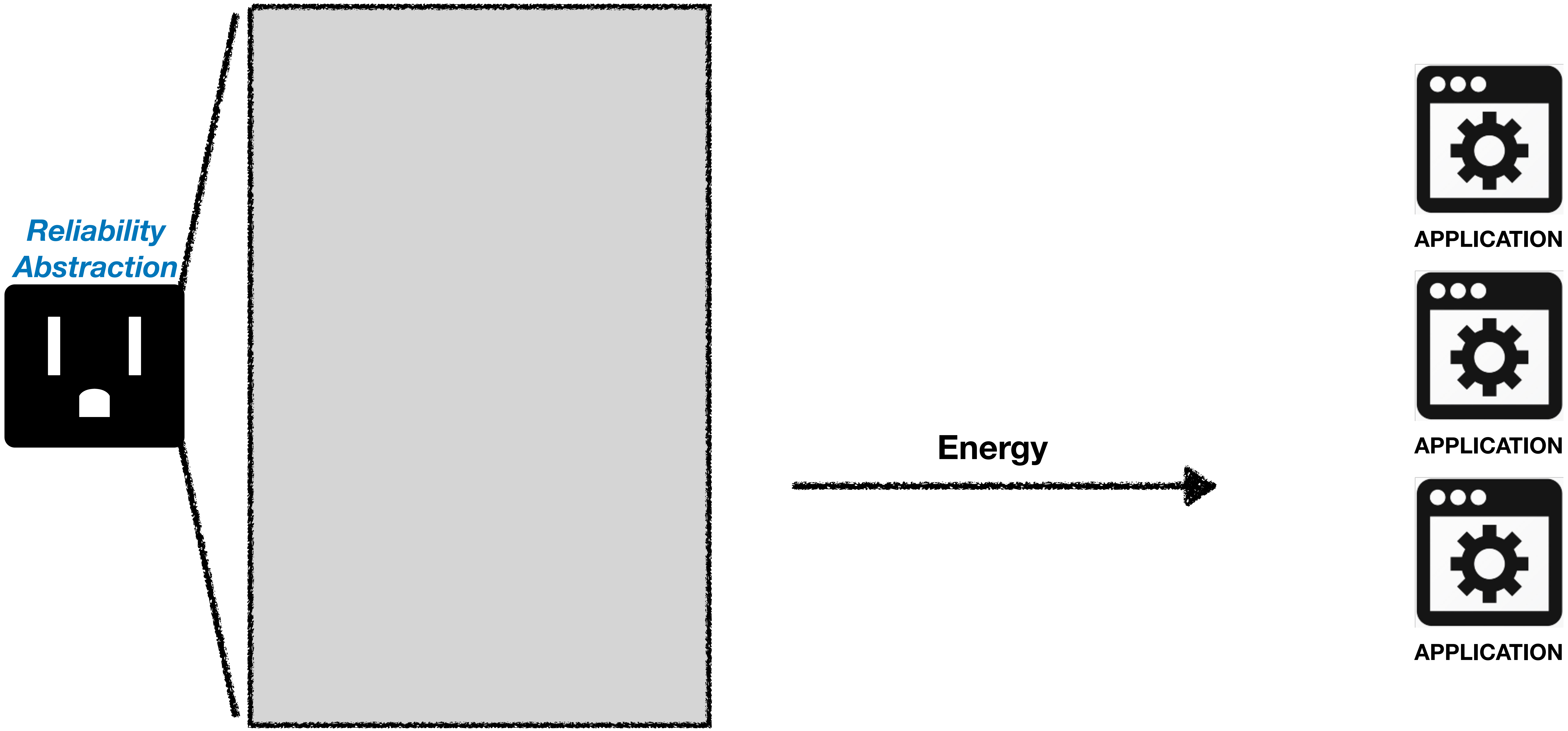
Collaborators:

* University of Massachusetts Amherst

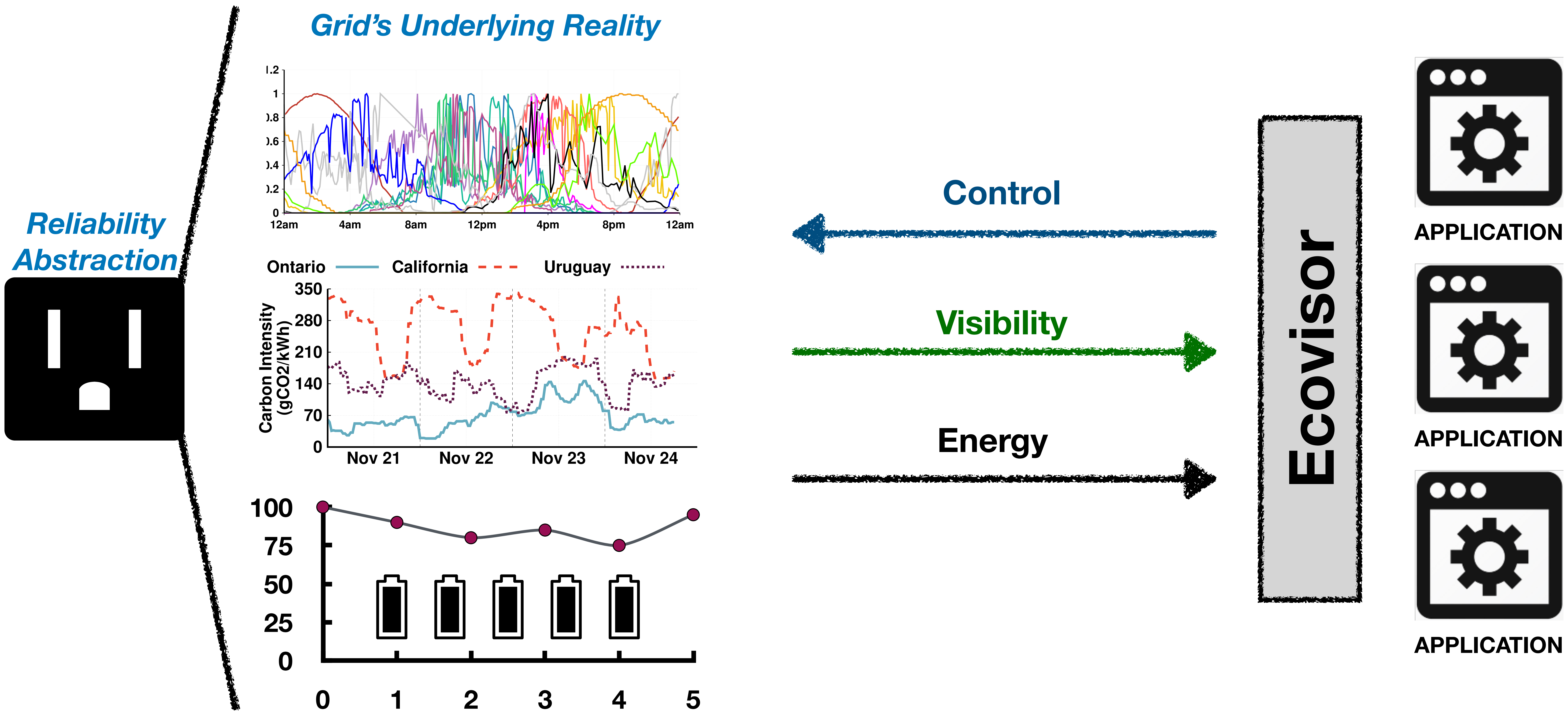
^ Worcester Polytechnic Institute (WPI)

^^ California Institute of Technology (Caltech)

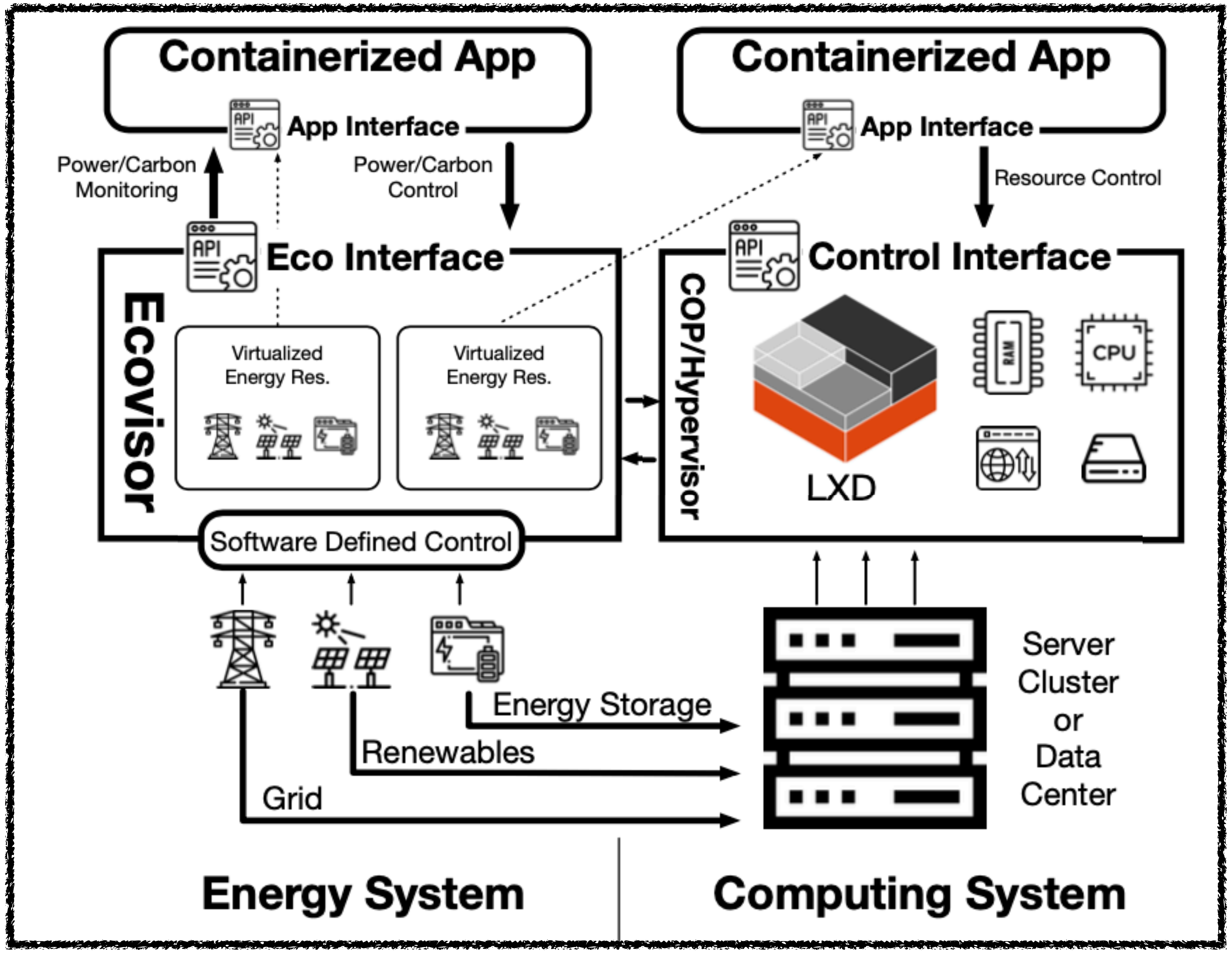
Ecovisor: A Virtual Energy System for Carbon-Efficient Applications



Ecovisor: A Virtual Energy System for Carbon-Efficient Applications



Ecovisor: Design and API



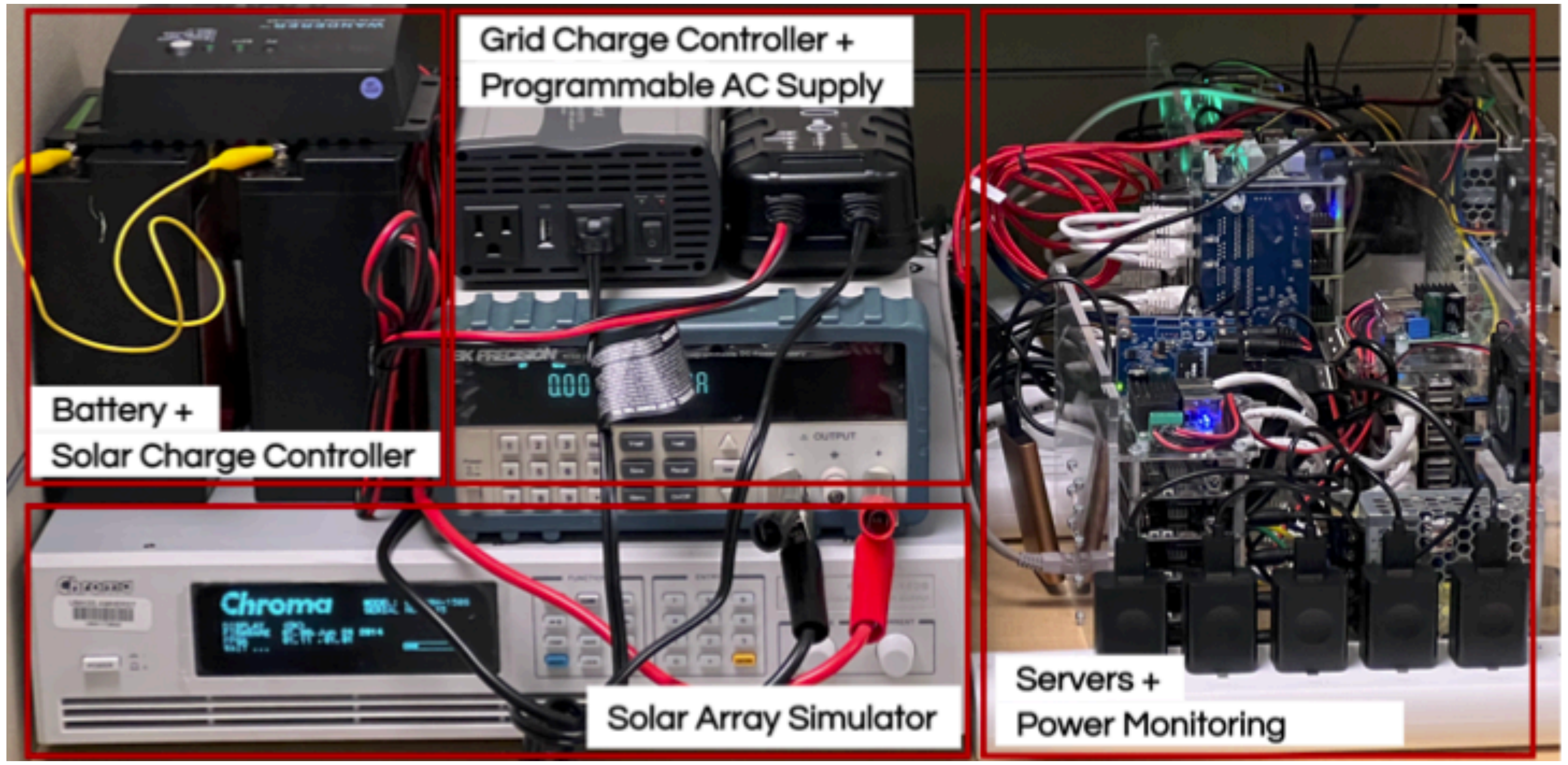
Function Name	Type	Input	Return Value	Description
set_container_powercap()	Setter	ContainerID, kW	N/A	Set a container's power cap
set_battery_charge_rate()	Setter	kW	N/A	Set battery charge rate until full
set_battery_max_discharge()	Setter	kW	N/A	Set max battery discharge rate
get_solar_power()	Getter	N/A	kW	Get virtual solar power output
get_grid_power()	Getter	N/A	kW	Get virtual grid power usage
get_grid_carbon()	Getter	N/A	$g \cdot CO_2/kW$	Get current grid carbon intensity
get_battery_discharge_rate()	Getter	N/A	kW	Get current rate of battery discharge
get_battery_charge_level()	Getter	N/A	kWh	Get energy stored in virtual battery
get_container_powercap()	Getter	ContainerID	kW	Get a container's power cap
get_container_power()	Getter	ContainerID	kW	Get a container's power usage
tick()	Notification	N/A	N/A	Invoked by ecovisor every Δt

Control Power Supply and Demand

Asynchronous Notifications

Get Energy System Information

Ecovisor: Prototype Implementation

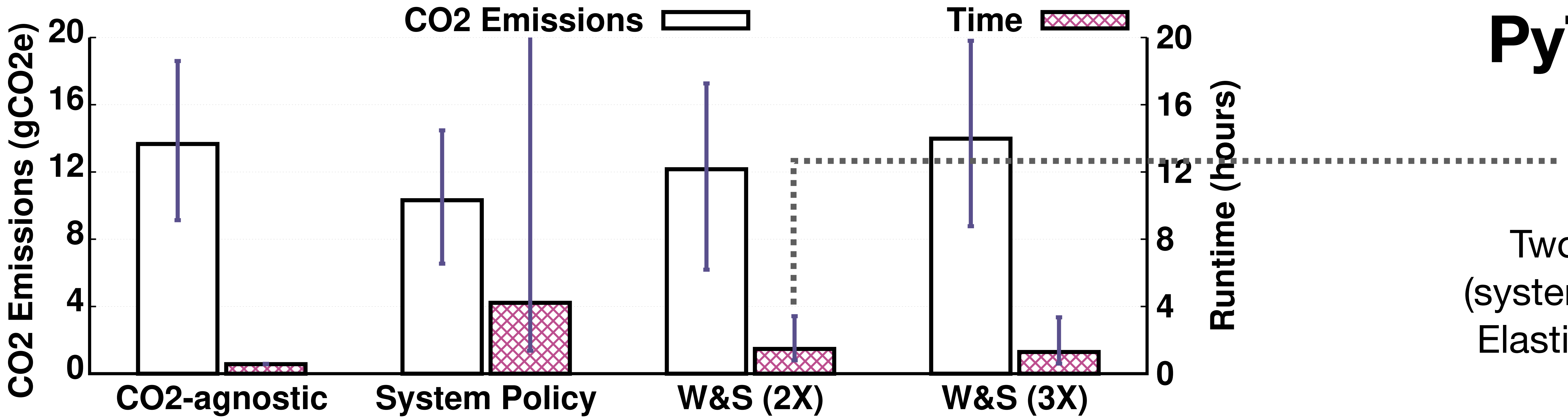


- **Software:** REST API
- **Hardware:** 60 Rock64 nodes

1. Reducing carbon (ML training, MPI)
2. Budgeting carbon (webserver)
3. Leveraging batteries (web server, Spark)
4. Leveraging solar (MPI, straggler)

Ecovisor: Optimizing Carbon/Performance Trade-off

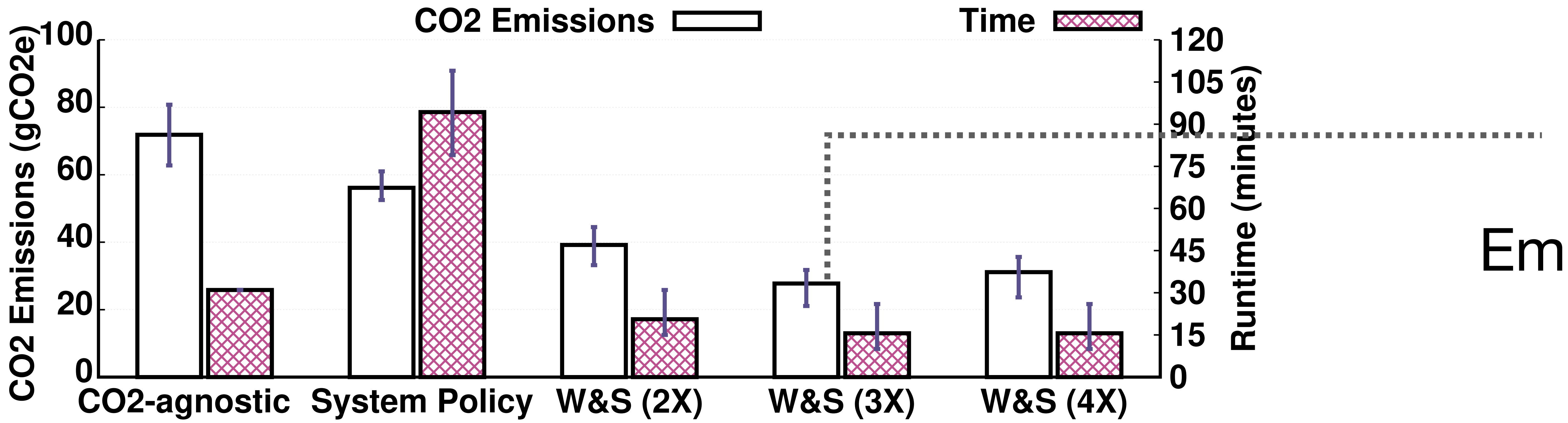
- **Evaluation objectives:** Demonstrate carbon savings, show applications should do optimizations.
- **Baseline:** (WaitAWhile - Middleware '21), **Proposed:** Application-specific (Wait&Scale) policy



PyTorch ML Training

Optimal Scale = 2X

Two follow-up papers, **CarbonScaler** (system) and **RORO** (theory), on leveraging Elasticity will appear at **SIGMETRICS'24**.

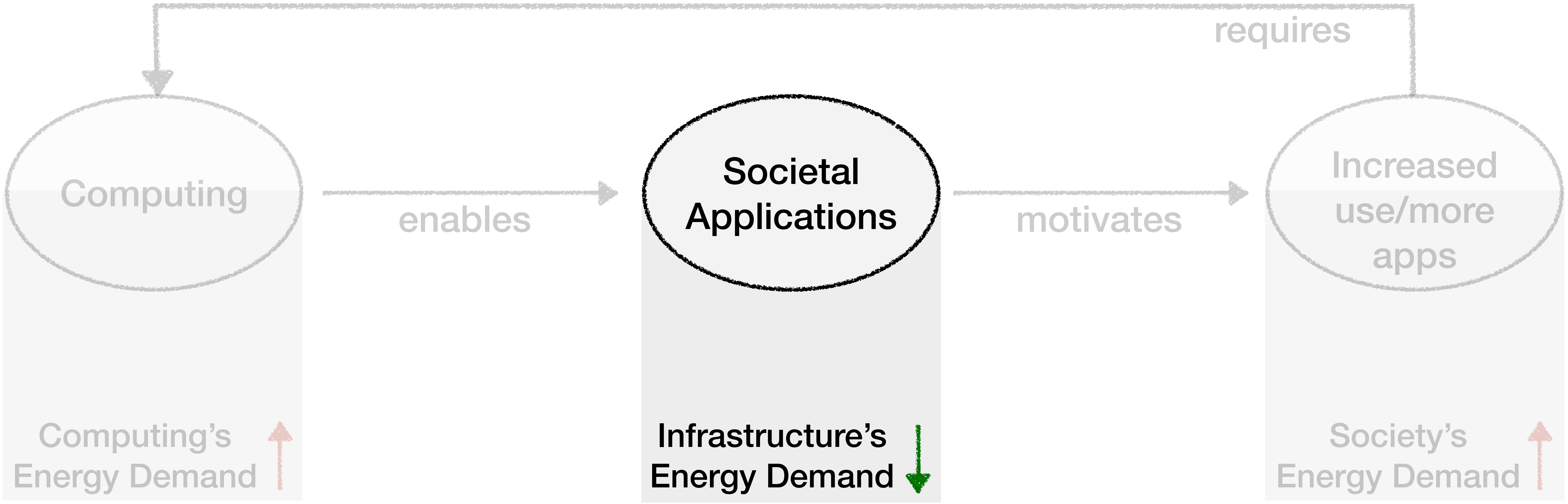


BLAST

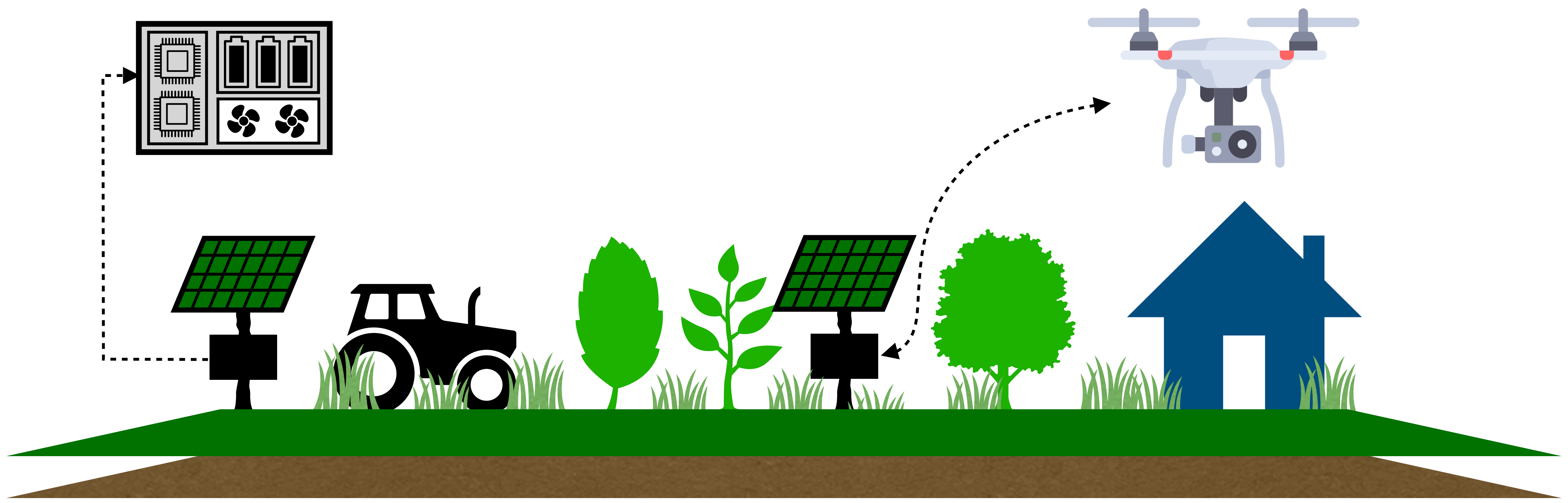
Optimal Scale = 3X

Embarrassingly parallel job.

Computing for Sustainability



Computing Use Cases



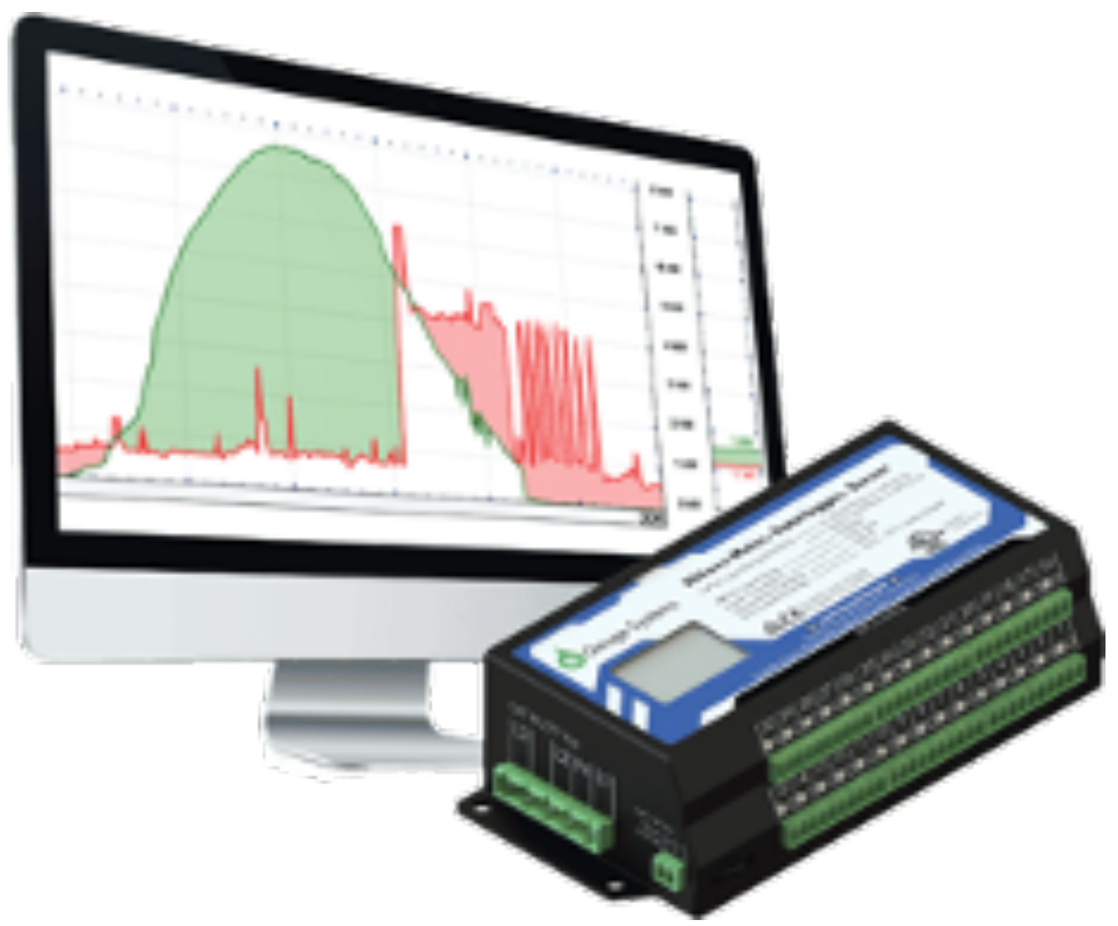
Improving Buildings and Transportation Sectors

- Building as an example of a distributed system
  - **Sense** monitor energy, temperature, occupancy etc.
  - **Analyze** data using computational tools.
  - **Control** lights, HVAC, doors to reduce energy usage.
- Transportation as an example of a distributed system
  - **Sense?**
  - **Analyze?**
  - **Control?**
- Agriculture as an example of computing use case
  - **Sense?**
  - **Analyze?**
  - **Control?**

Building Monitoring

- Power metering at different levels
    - Outlet-level monitoring
    - Meter-level monitoring

Wemo smart plug



eGauge meter
with interface

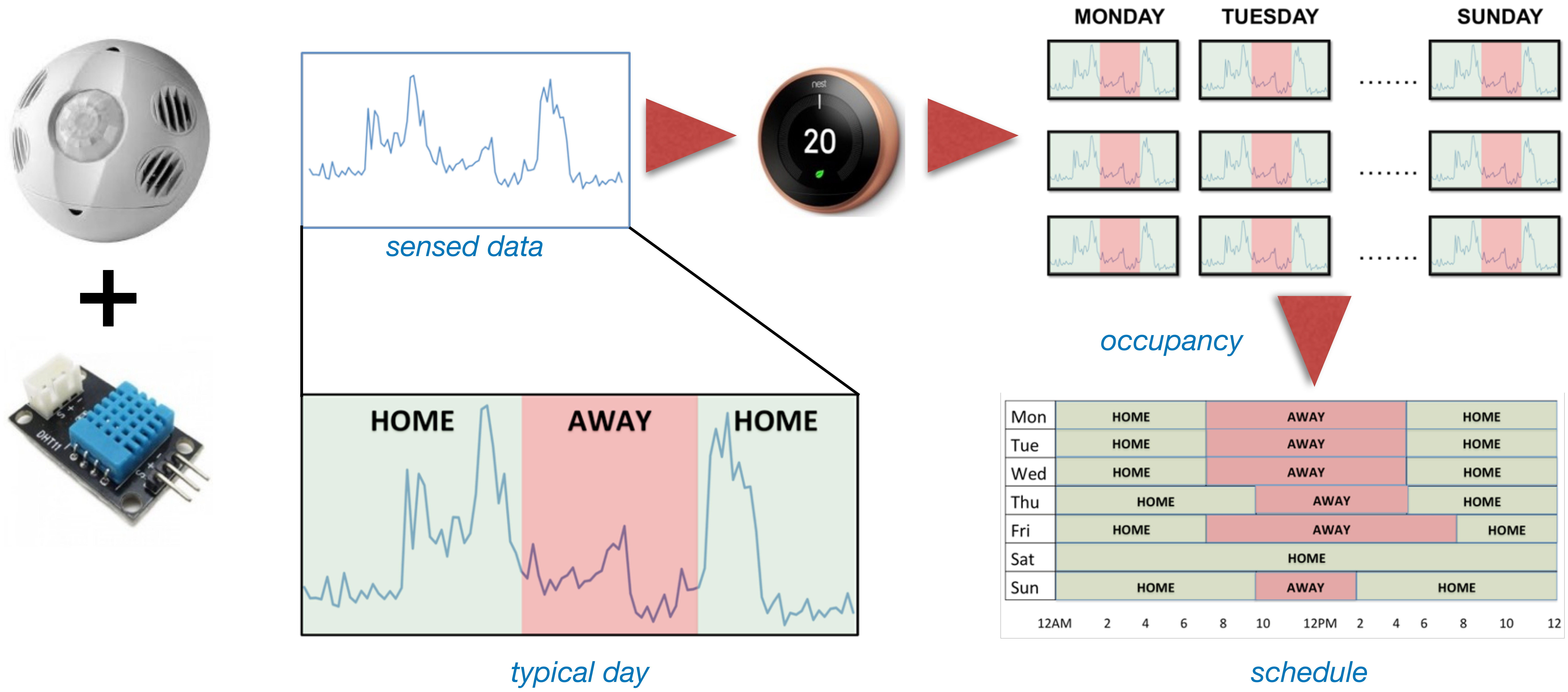
smart meter

Analyzing the data

- Energy monitors / sensors provide real-time usage data
  - Building monitoring systems (BMS) data from office / commercial buildings

- Modeling, Analytics and Predictions
  - Use statistical techniques, machine learning and modeling to gain deep insights
  - Which homes have inefficient furnaces, heaters, dryers?
  - Are you wasting energy in your home?
  - Is an office building’s AC schedule aligned with occupancy patterns?
  - When will the aggregate load or transmission load peak?

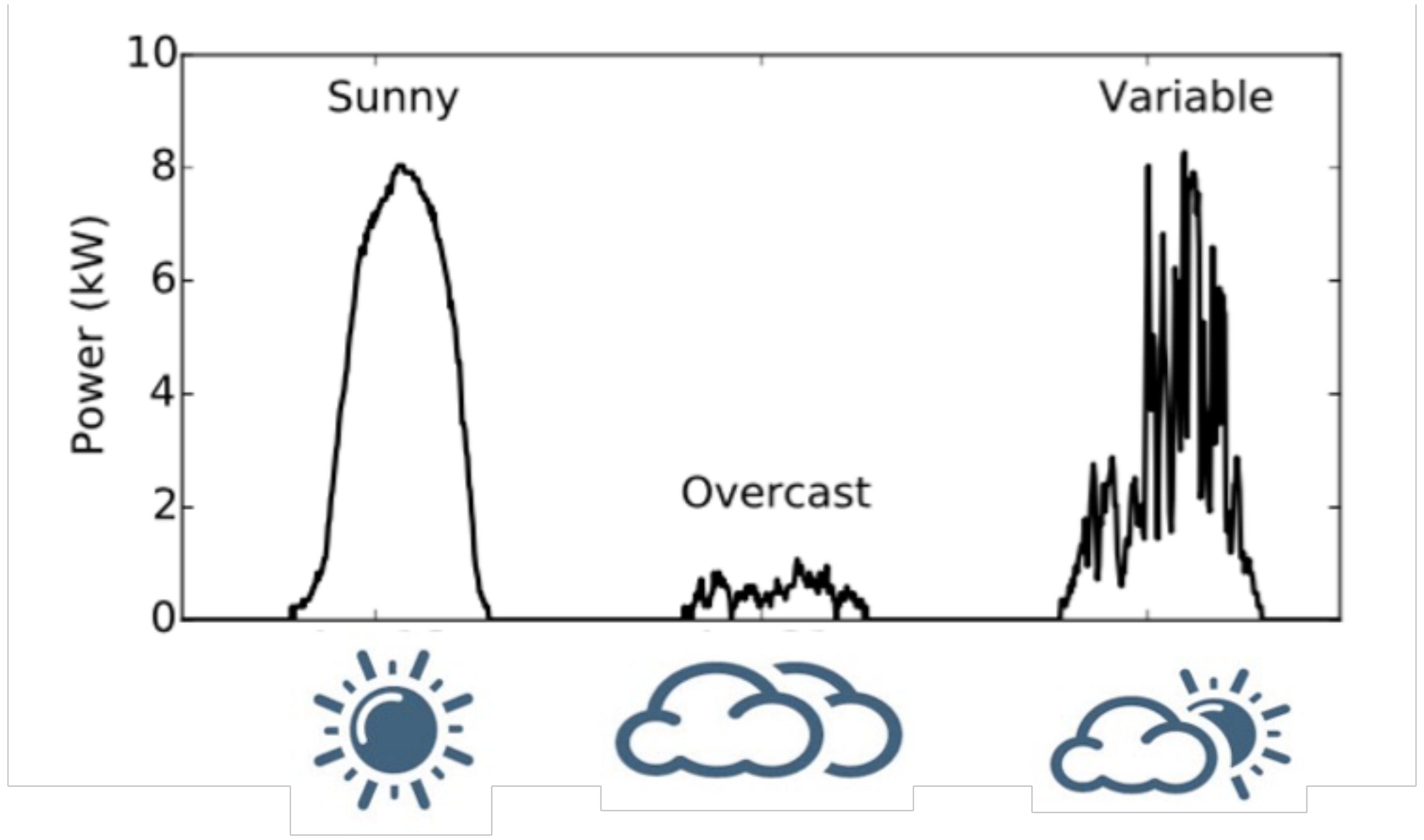
Reduce Energy Use —> Learning Thermostat



Use Low Carbon Energy —> Use Solar Power

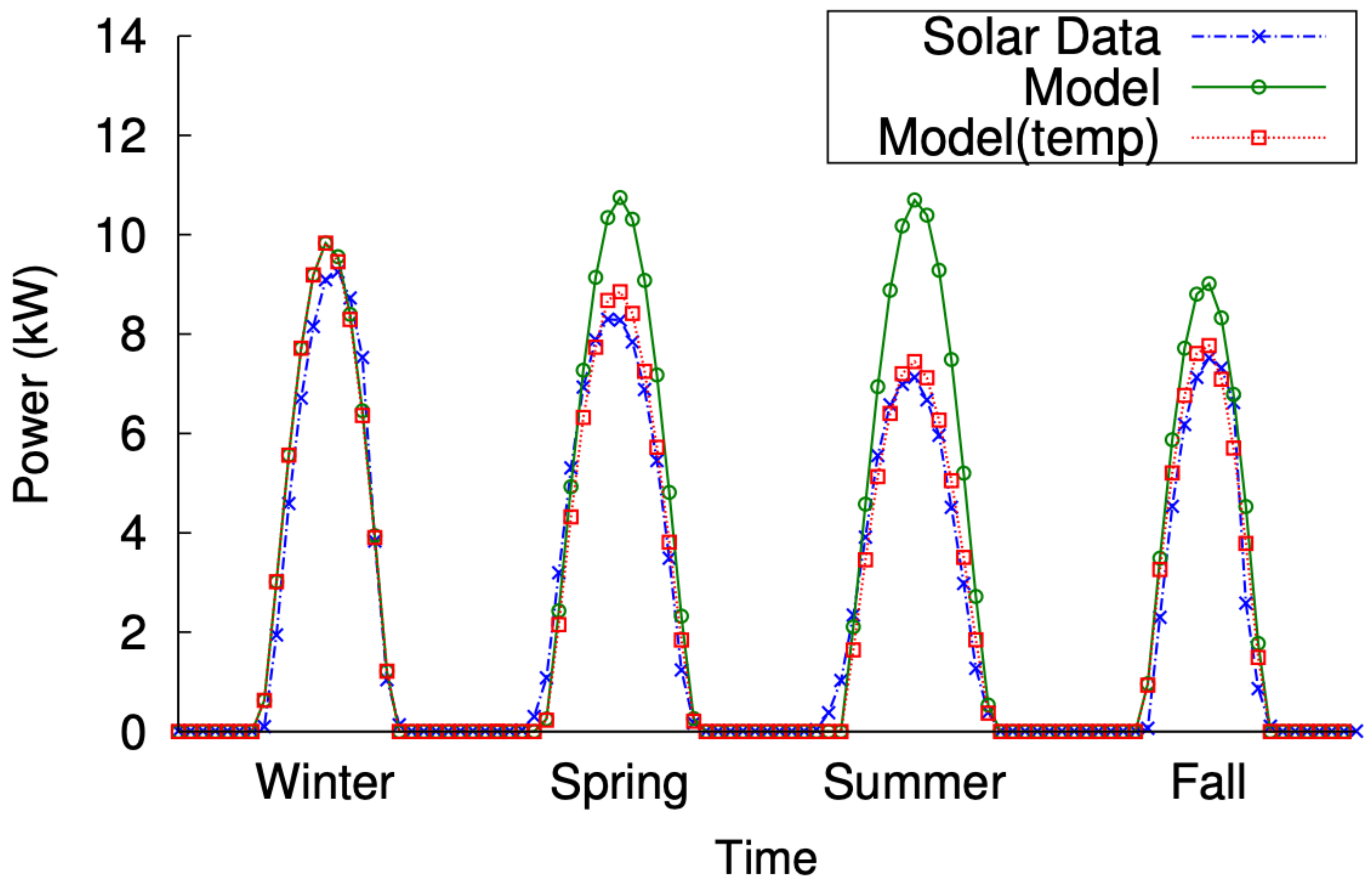
- Significant growth in renewable energy adoption
  - Roof top wind turbines, solar PV, solar thermal (water heating)

- Highly intermittent
  - Impacted by cloud cover, temperature, environmental variables



Forecasting Solar Energy

- Predictive analytics to model and forecast solar energy generation
    - Use machine learning and NWS weather forecasts to predict solar generation



- Better forecasts of near-term generation; “Sunny load” scheduling

Use Case - EV Charging

- Solar panels installed in parking lots, rest areas, paid garages
  - Possible use case in offices and car rental services

- Assumptions
  - Arrival/departure times for EVs
  - Accurate solar predictions

- Intelligent charging
  - When to charge?
  - Which EV to charge?
  - How much to charge?

Climate and Sustainability Implications of Generative AI

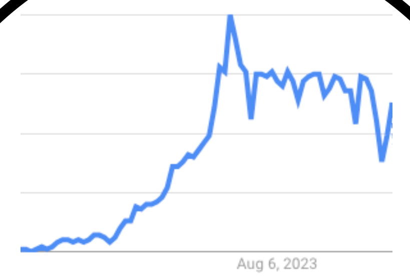
Noman Bashir[1], Priya L. Donti[2,3], James Cuff[4], Sydney Sroka[1], Marija Ilic[2,3], Vivienne Sze[4,5,6,7], Christina Delimitrou[7], Elsa A. Olivetti[1,8]

[1] MIT Climate & Sustainability Consortium (MCSC),
[2] MIT Electrical Engineering and Computer Science (EECS),
[3] MIT Laboratory for Information & Decision Systems (LIDS),
[4] MIT Office of Research Computing & Data (ORCD),
[5] MIT Research Lab of Electronics (RLE),
[6] MIT Microsystems Technology Laboratories (MTL),
[7] MIT Computer Science & Artificial Laboratory (CSAIL).
[8] MIT Materials Science & Engineering (DMSE)

Unfettered Growth and Its Key Drivers

ChatGPT 1 million users in 5 days

Only 15% of the users are from US



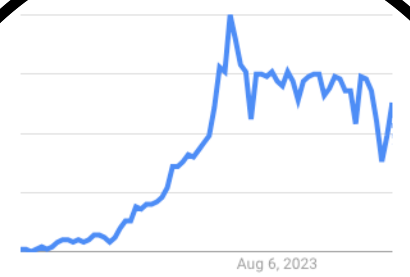
Interest in Gen-AI

perceived benefits

Unfettered Growth and Its Key Drivers

ChatGPT 1 million users in 5 days

Only 5% of the users are from US

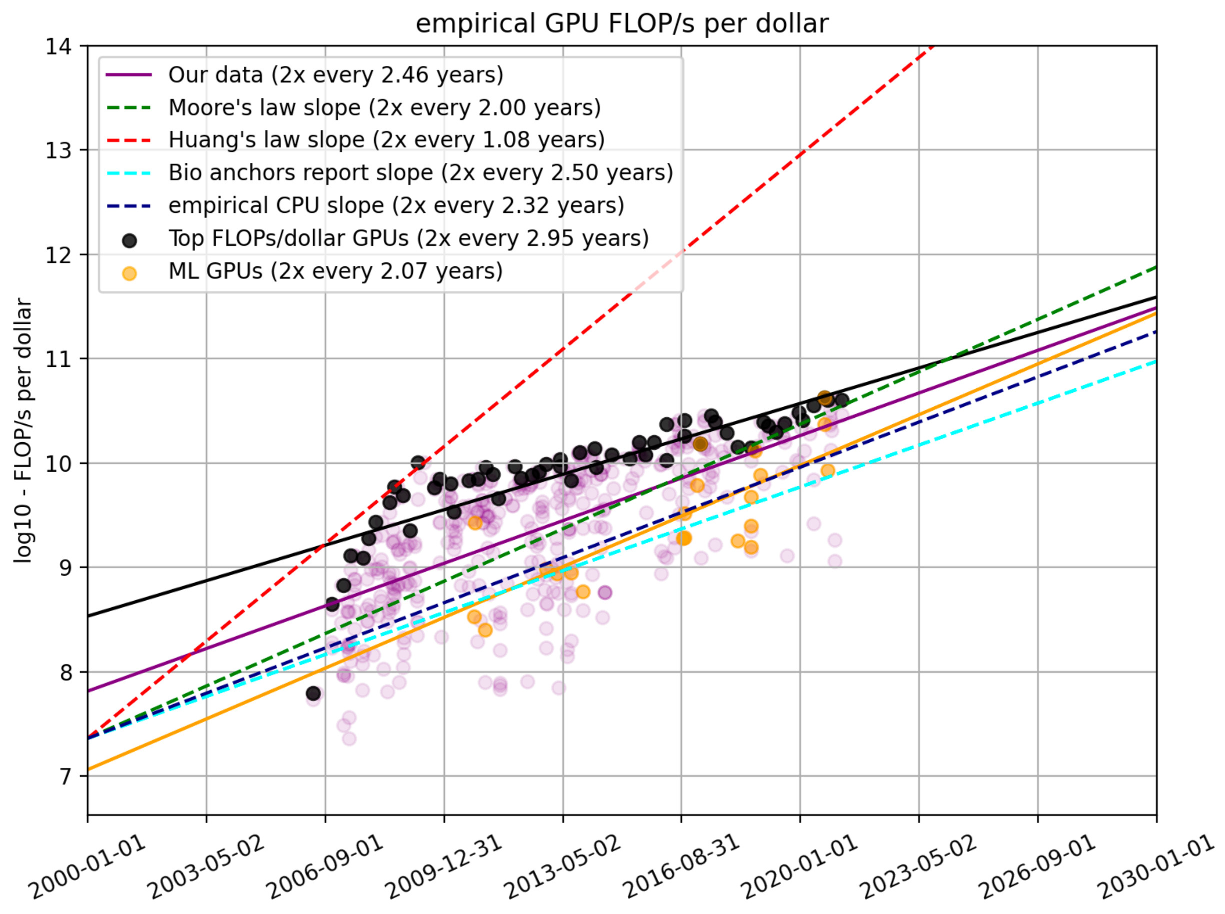


Interest in Gen-AI

perceived benefits

consolidation of AI capabilities

lack of regulatory oversight

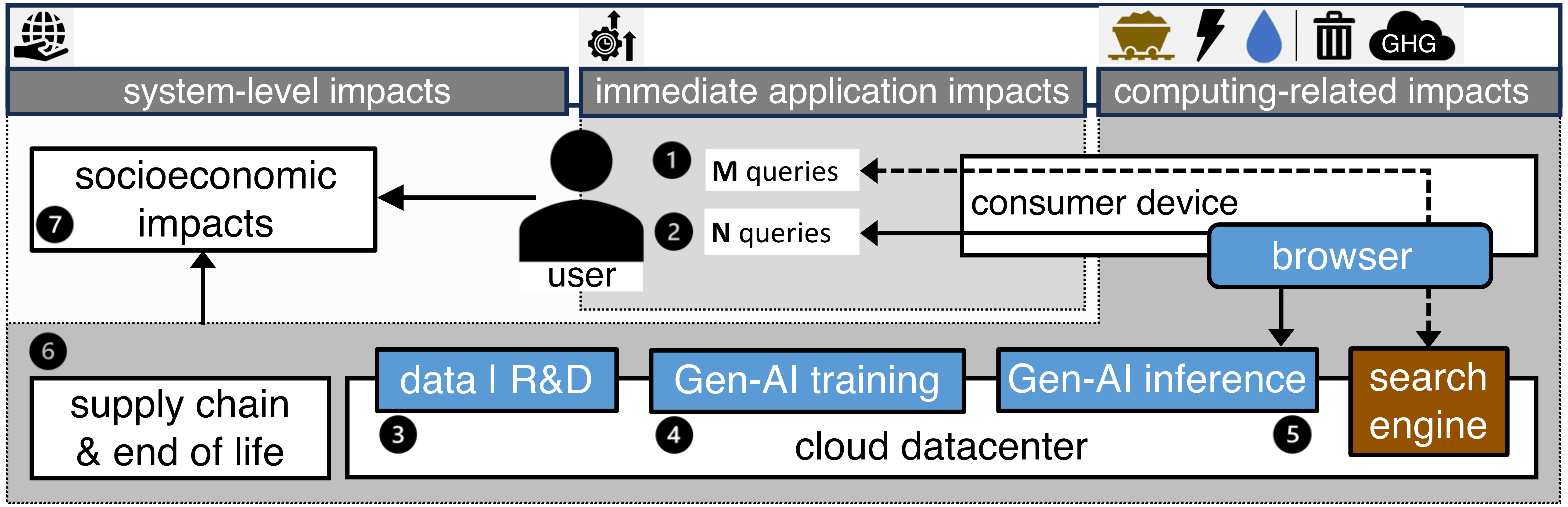


efficiency improvements

Need for Comparative Benefit-cost Evaluation Capability

- Scope
    - E.g., a search query.
- Boundaries
    - Geographical: A given region or a data center.
    - Temporal: A given window of time.
    - Conceptual: A search query.
- Baselines and scenarios
    - A standard Google search as a baseline.
    - Various GPT models as scenarios.
- Metrics and data
    - Energy usage, GHG emissions, water usage, and raw material.

Illustrative Example: Generative AI-based Search



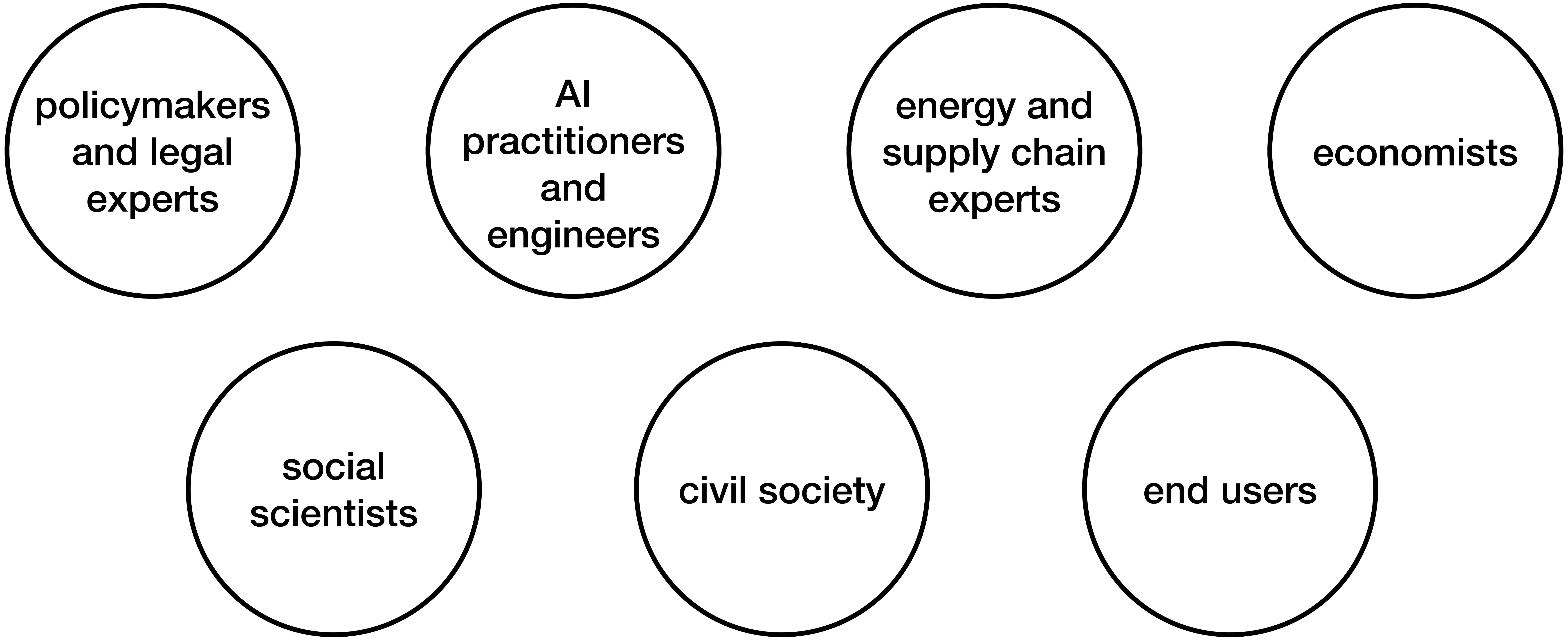
1. **Baseline:** User needs **M** queries.
2. **Gen-AI:** User needs **N** queries.
3. **Baseline & Gen-AI:** Both incur costs during data processing and R&D phase.
4. **Gen-AI:**, Model training is an additional cost.
5. **Baseline & Gen-AI:** Both incur per-query costs, which may differ.
6. **Baseline & Gen-AI:** Both incur costs during supply chain and end-of-life phases.
7. **Baseline & Gen-AI:** User's actions have system-level socioeconomic impacts.

The **computing-related** costs include raw material usage, energy consumption, waste generation, and water use.

The **immediate application** impacts include the reduced time spent on search and quality of response.

The **system-level** impacts include broader socioeconomic impacts computing as well as user using the Gen-AI for search.

Stakeholder Engagement for Responsible Development in Gen-AI



Leveraging Benefit-cost Evaluation Framework

- Monitoring the evolution of Gen-AI as a sector
- Identifying opportunities to improve benefit-cost ratio
- Facilitating eco-economic decoupling and constrained growth

Summary

- Sustainable Computing
  - Demand for computing is growing
  - Need to serve the demand sustainably
  - Energy efficiency gains reducing
  - Computing has unique advantages
  - Try to optimize computing’s carbon efficiency
  - Reduce operational as well as emobodied carbon

- Computing for Sustainability
  - Leverage computing to reduce energy consumption
  - Leverage computing to enhance use of low carbon energy