University of Massachusetts Amherst

Data Centers, AI, and Efficiency: A Systems Perspective

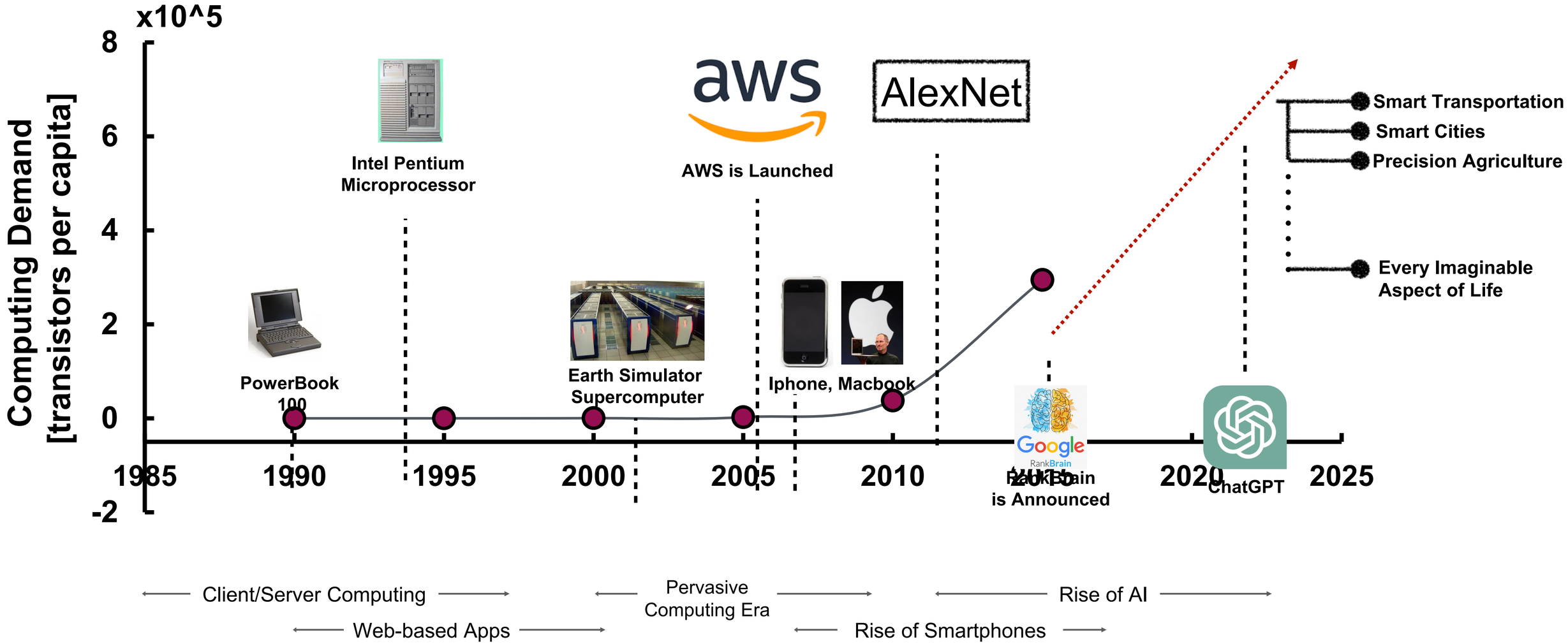
Prashant Shenoy

University of Massachusetts

Dec 8, 2025, NSF Compute-Energy Nexus Workshop

Computing's Demand is Growing Exponentially

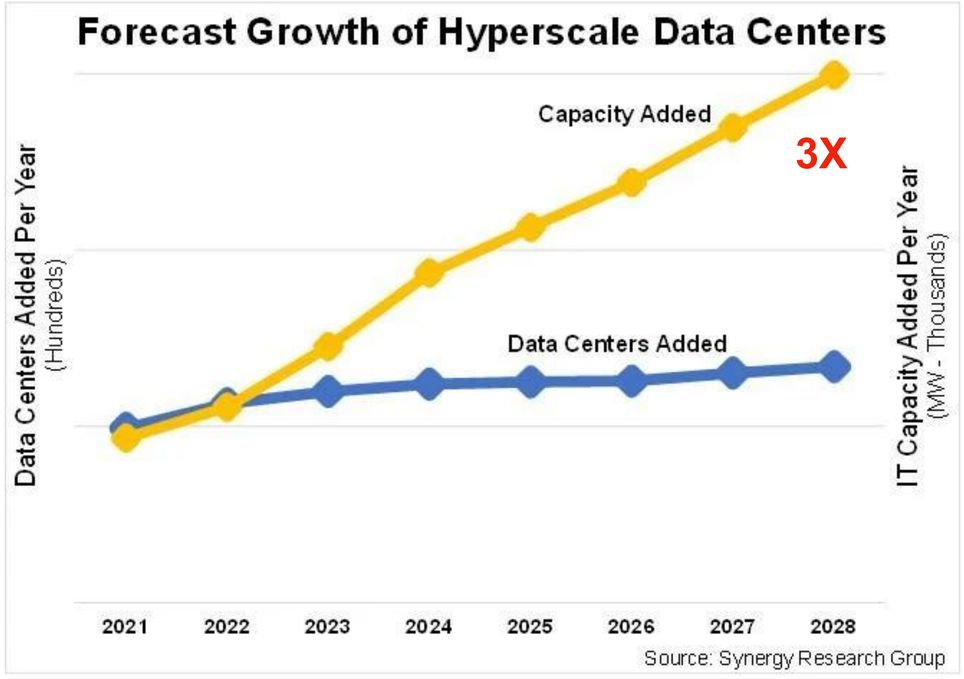
- Defining trend of our time: internet, mobile, cloud, and now AI



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

Data Centers Growth and Impact on Energy Use

- LBL Study on data center energy growth in the AI era
- Previously: Energy use grew more slowly due to energy/PUE optimizations
- Future: data centers will comprise 6 - 12% of energy use by 2028



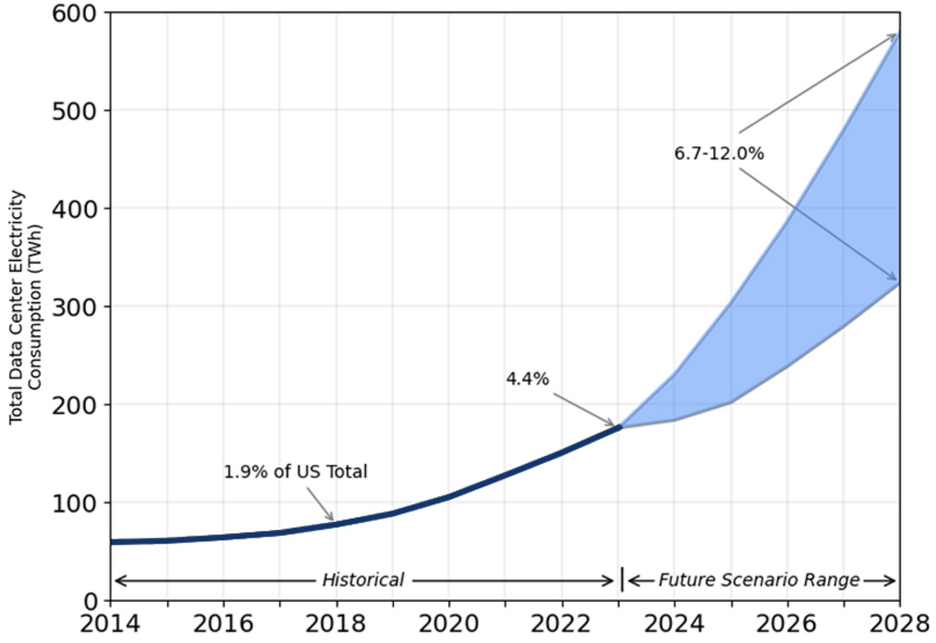


Fig courtesy: LBL 2024

Can the Electric Grid keep up?

- Grid capacity growth is struggling to keep up with data center growth.

US electric grids under pressure from energy-hungry data centers are changing strategy

Assessing the AI Data Center Sector Amid Grid Strain and Local Backlash

Rhys Northwood • Friday, Aug 29, 2025 2:44 pm ET 3min read

America's largest power grid is struggling to meet demand from AI

By Laila Kearney

July 9, 2025 10:25 AM UTC · Updated July 9, 2025

Aug 29, 2025 - Energy & Climate

America's power bills surge as AI strains an aging grid

Industry Conflicts

- Computing Industry Perspective

“We spent $500M for building this data center and need to utilize it fully to recoup investment”

- Grid Industry Perspective

“Large data center loads will destabilize the grid”

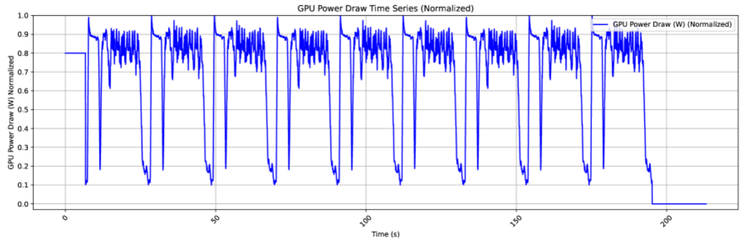


Fig courtesy: Chouske 2025

Need for Grid-friendly Computing Systems

Alberta's power grid 'cannot possibly connect' all proposed data centres, system operator says

Data center grid restrictions

Limit on power ramp up rates

Oscillations restricted

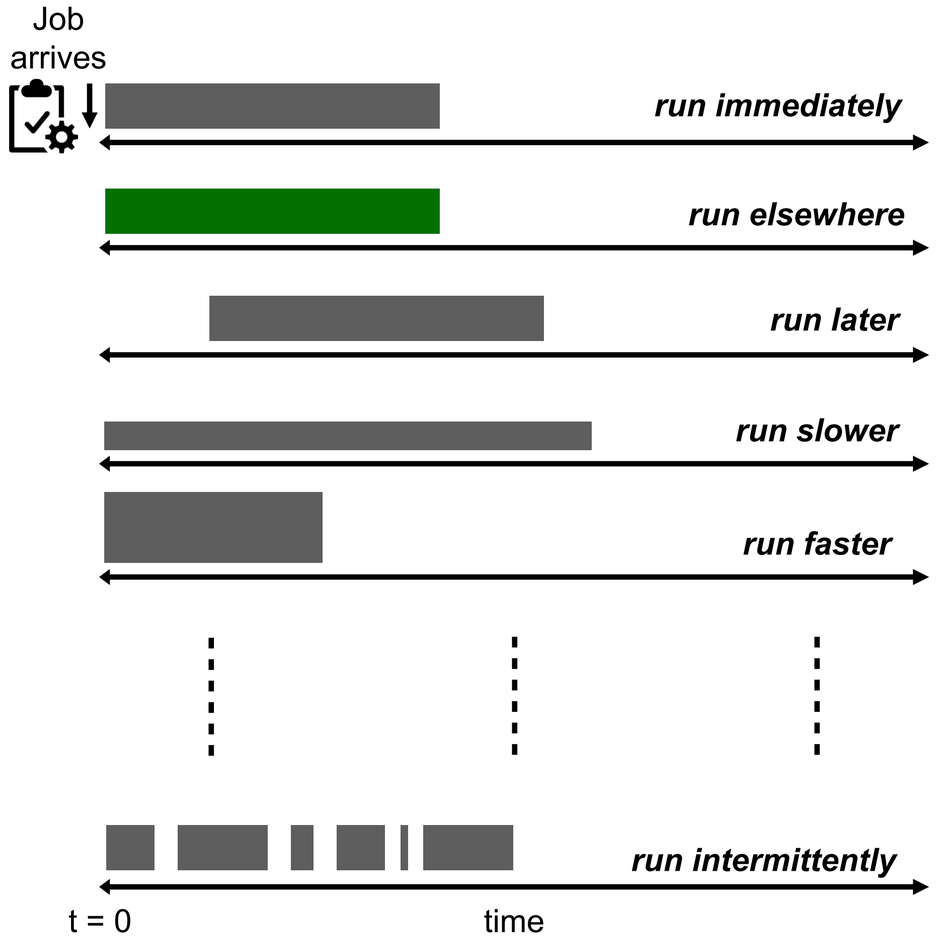
Must survive grid faults

Load shedding built-in

Data centers and computing systems need to become grid-friendly

What does it mean?

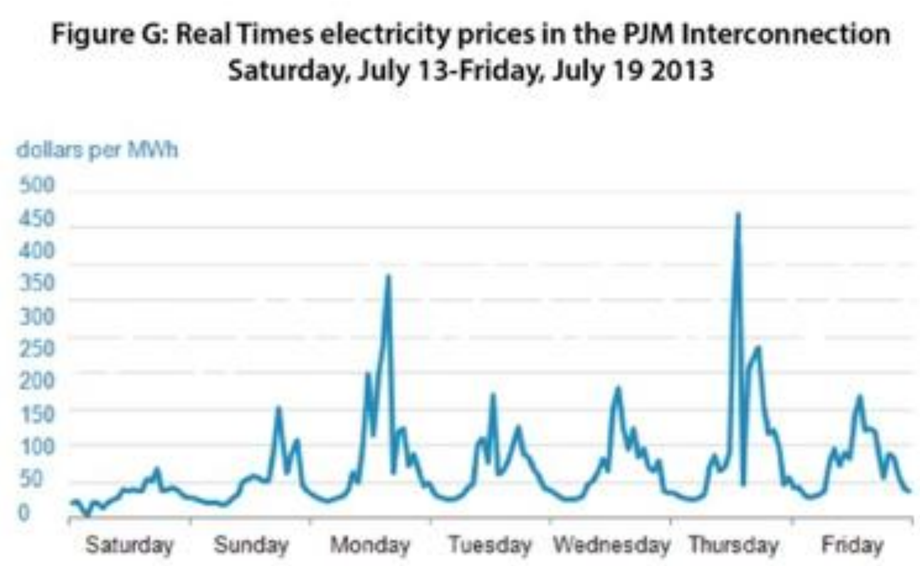
Computing workloads are uniquely flexible

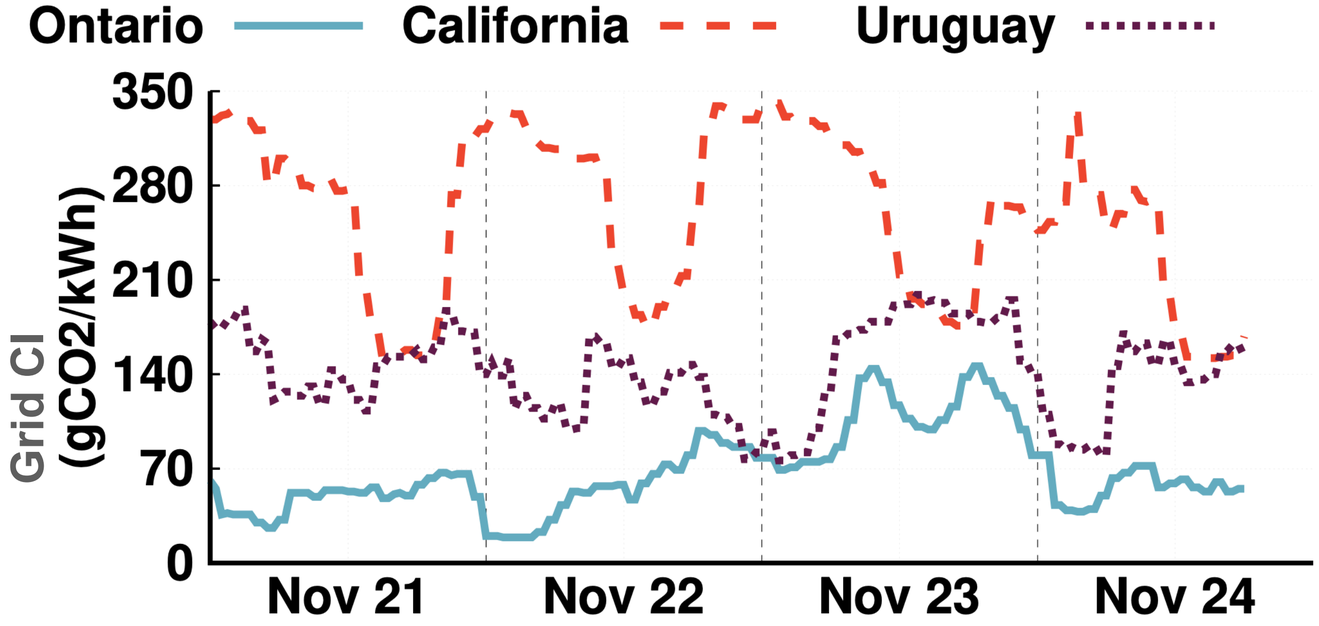


How can we leverage workload flexibility to make computing systems grid-friendly?

Potential solution - Workload Shifting

- Grid dynamics: availability and cost of electricity varies across regions and time
  - Temporal and spatial variations





Shift flexible workloads across time & regions based on grid dynamics

Time shifting via Suspend-Resume

- Time shifting: **suspend** when grid has high demand (cost), **resume** when demand (cost) drops
- Suspend-resume is grid-friendly but incurs **large increase** in completion time

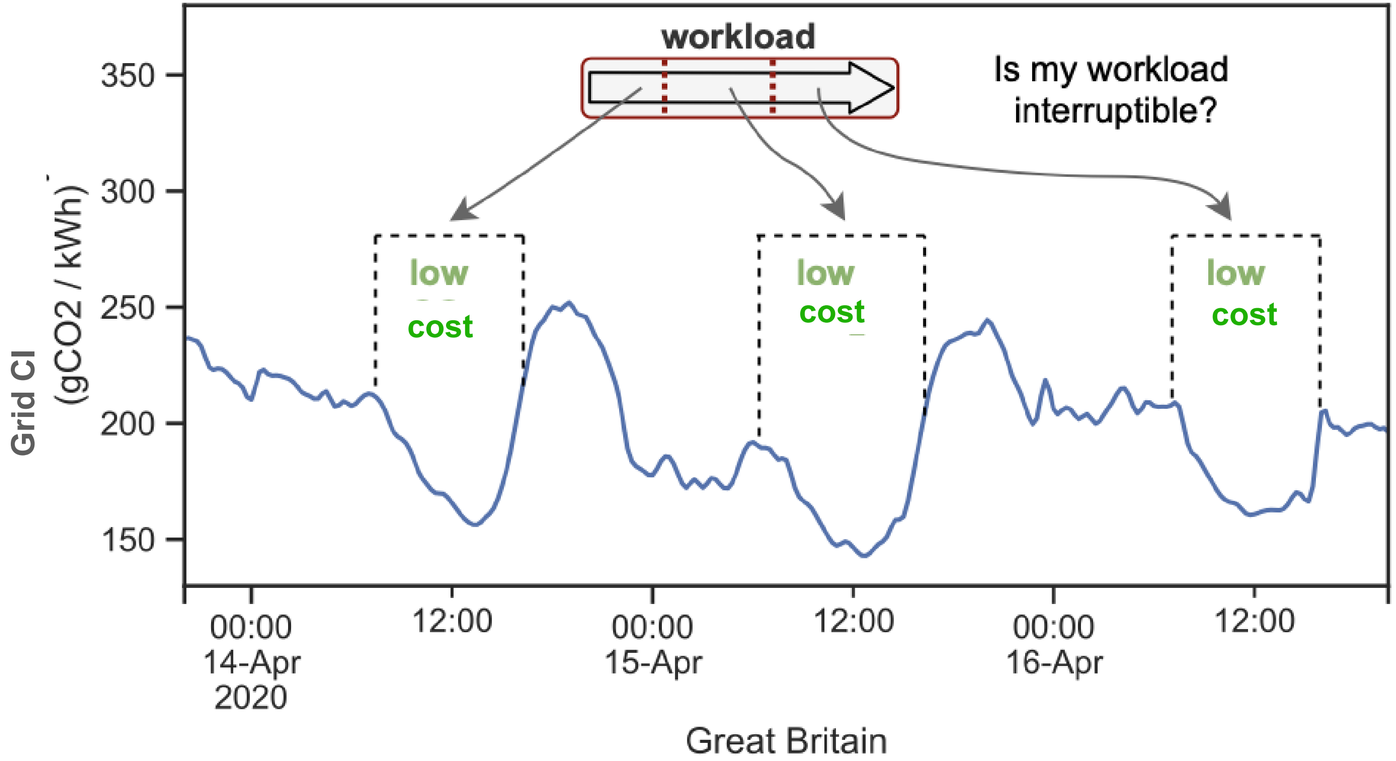
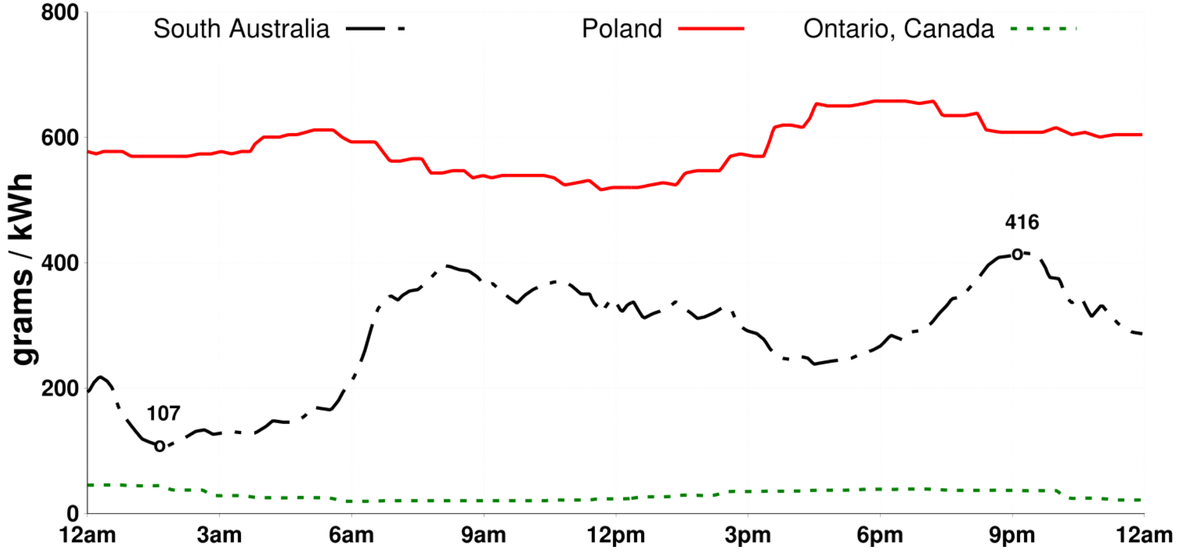


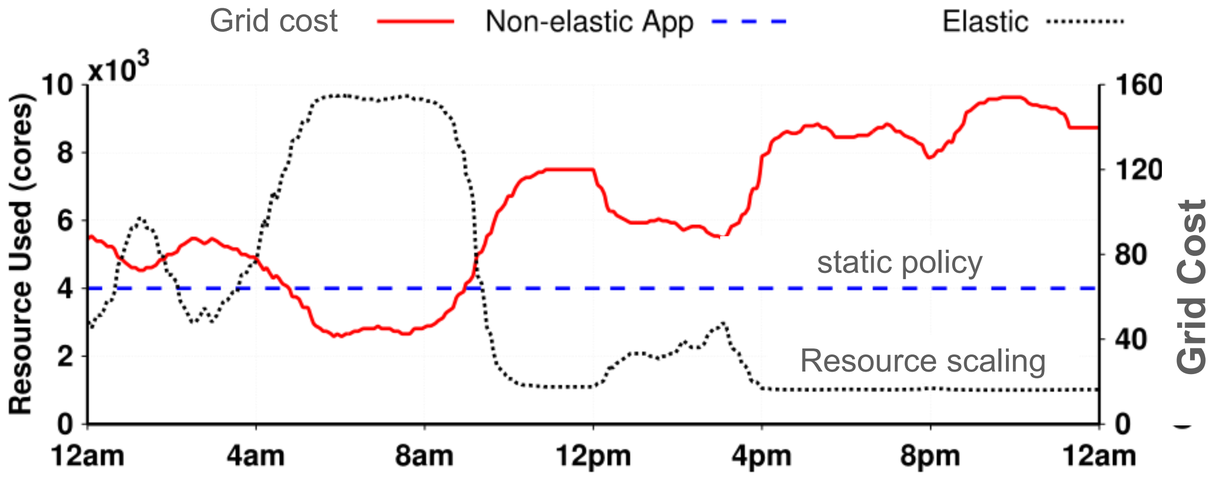
Fig courtesy: Let's Wait a While [Wiesner'21]

Grid-friendly Machine Learning Training

- Exploit elastic nature of machine learning training
- **Resource Scaling:** scale resources up and down to match grid fluctuations



Cost of electricity

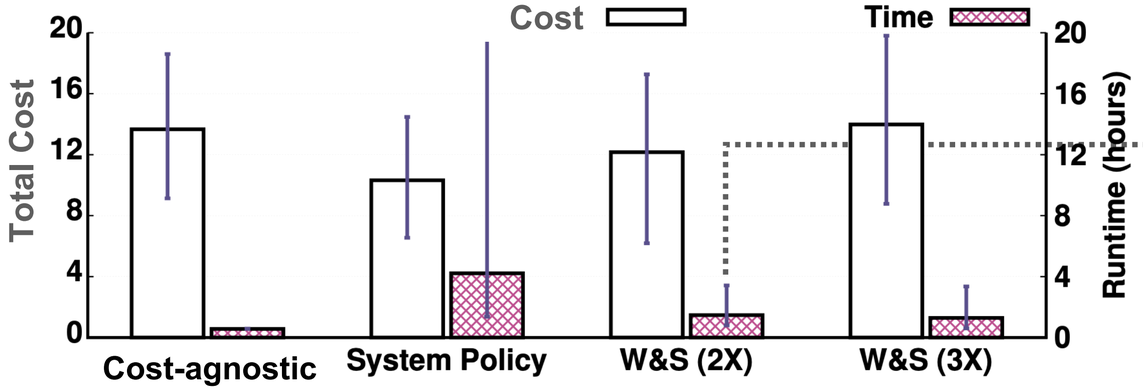


Scale up resources when cost is low

Avoids delays in completion time

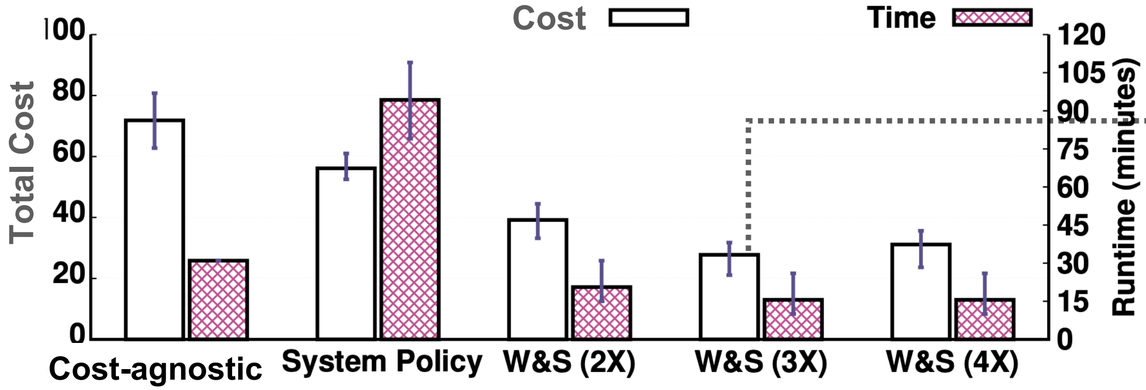
Cost-aware Resource Scaling

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



PyTorch ML Training

Optimal Scale = 2X



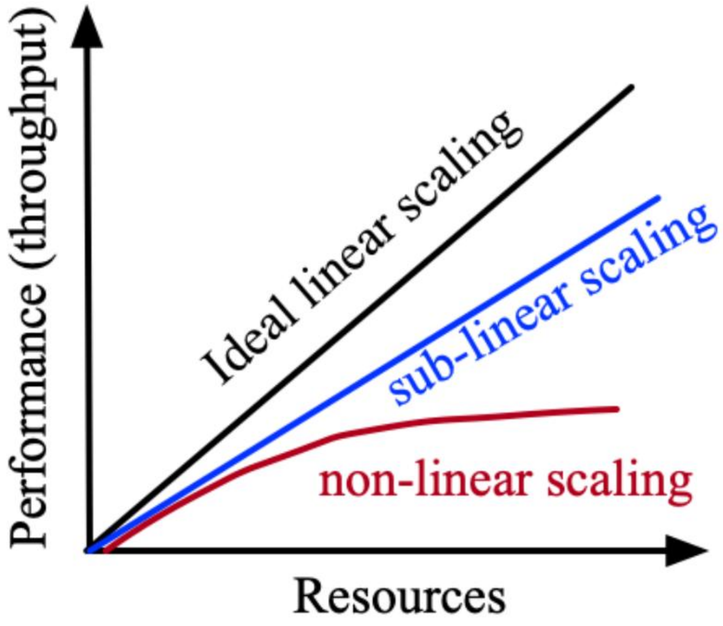
BLAST

Optimal Scale = 3X

Embarrassingly parallel job.

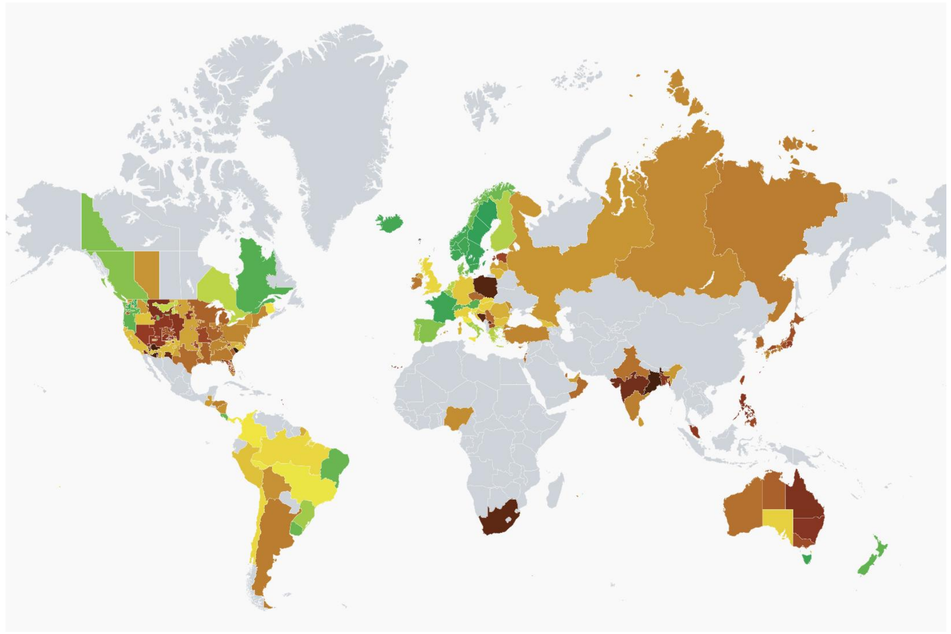
Challenges in Continuous Scaling

- Distributed cloud applications rarely scale linearly
  - Sub-linear or non-linear scaling common due to hardware/software bottlenecks
- Scaling up reduces efficiency, scaling down reduces performance



How much to scale, when, and where?

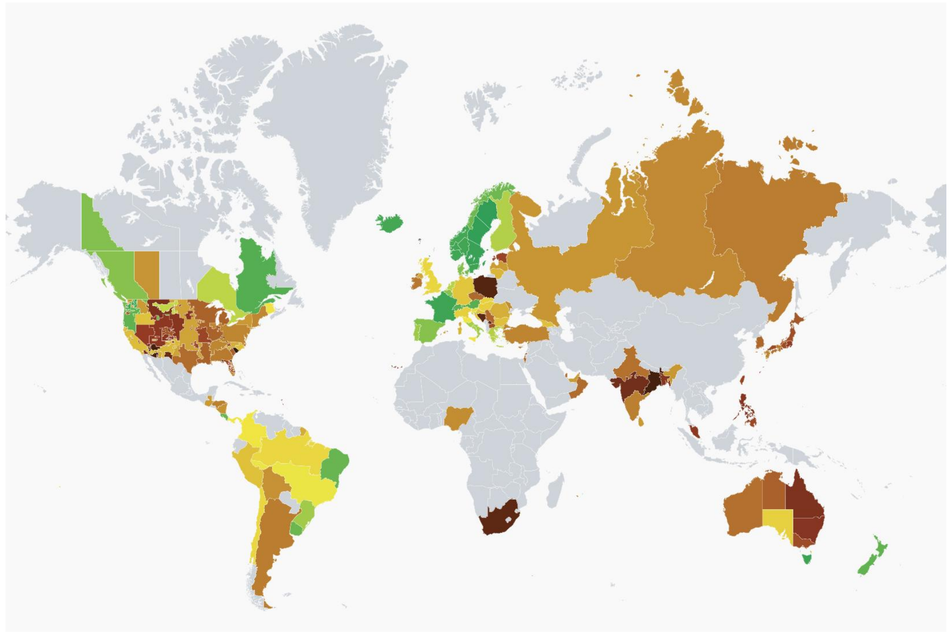
Spatial Workload Shifting



Energy demand, cost, and supply mix change across regions

Cloud data centers are distributed across continents.

Spatial Workload Shifting



Energy demand and supply mix change across regions.

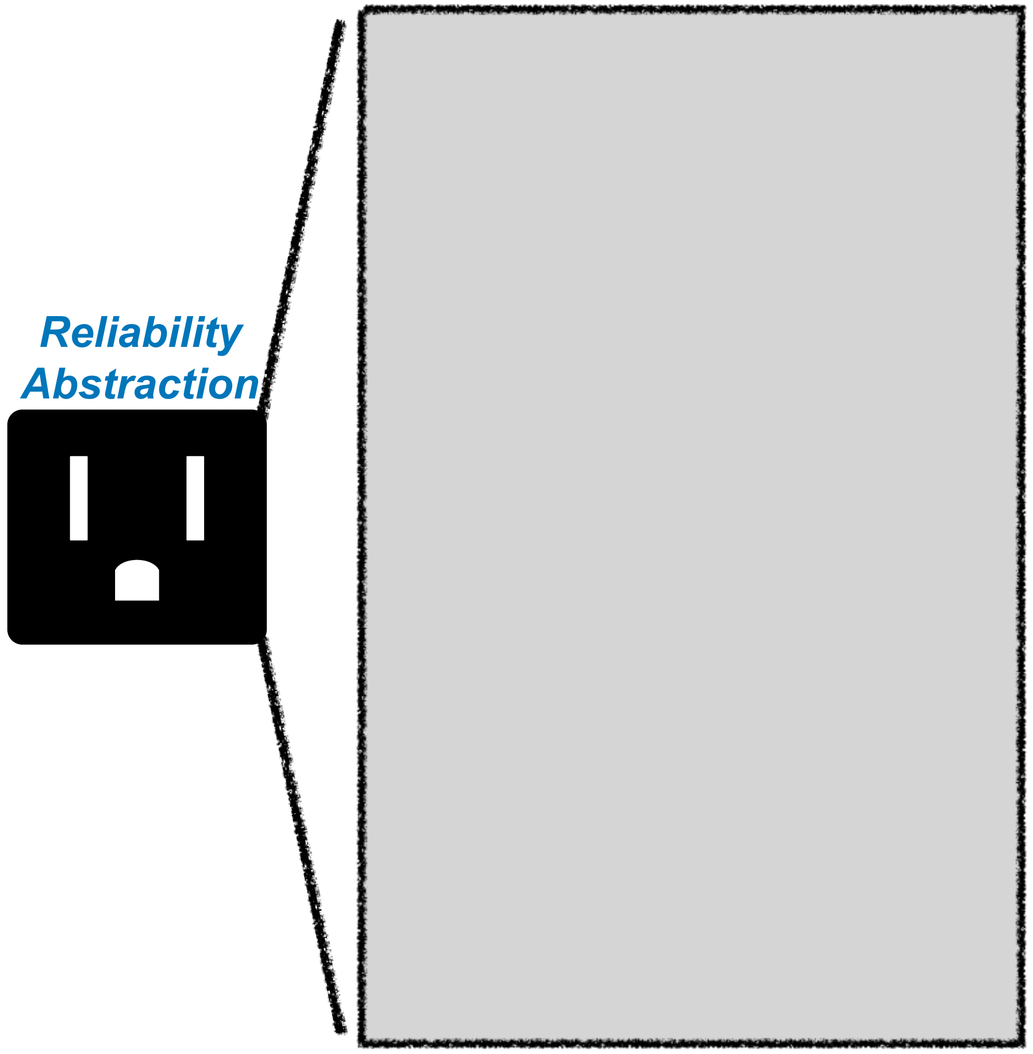
Cloud data centers are distributed across continents.

Spatial workload shifting at large geographic scales is grid-friendly but incurs large network costs (data/state movement, latency).

Open Challenges in Grid-friendly System Design

- How can grid dynamics be exposed to computer systems (data centers)?
- What are tradeoffs in designing grid-friendly computing systems?

Ecovisor Energy Hypervisor



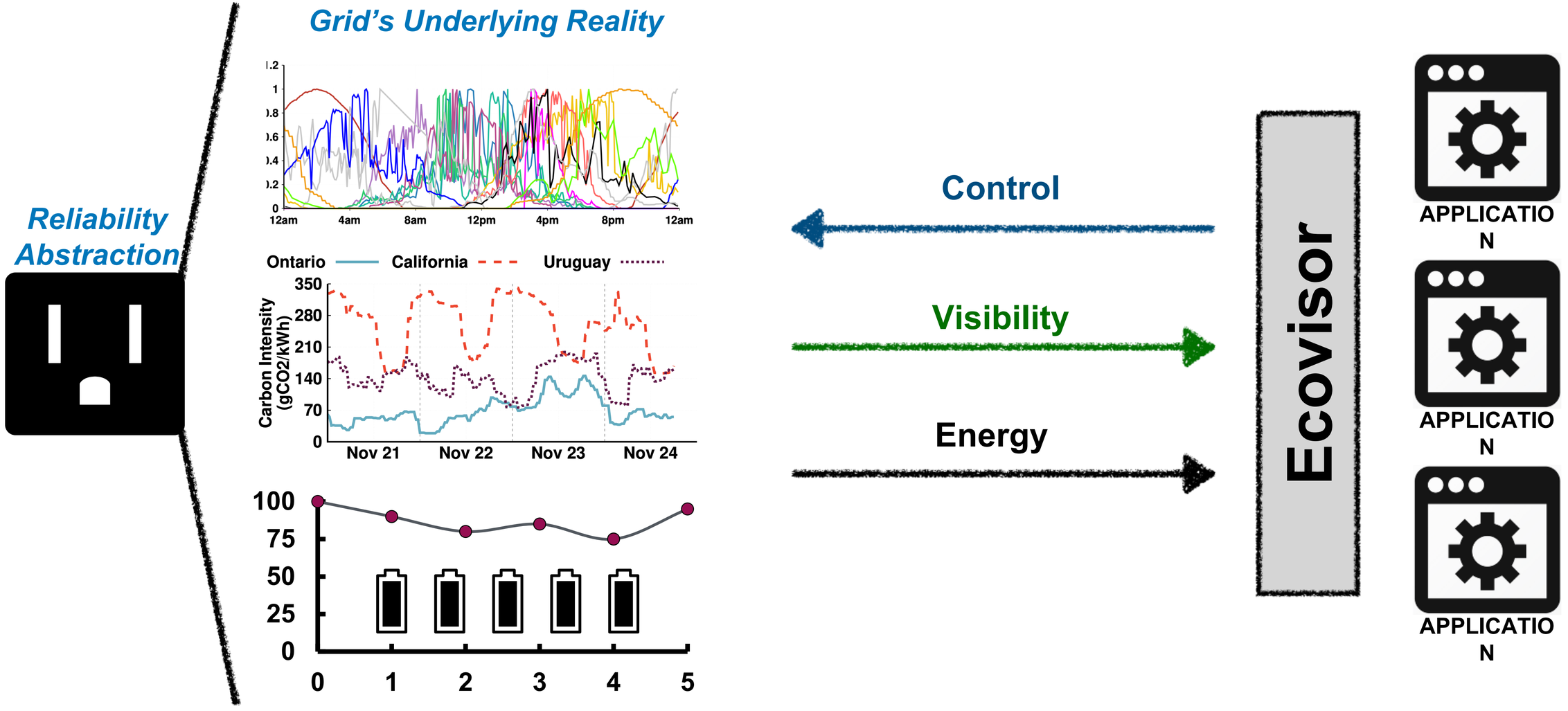
Energy

APPLICATIO
N

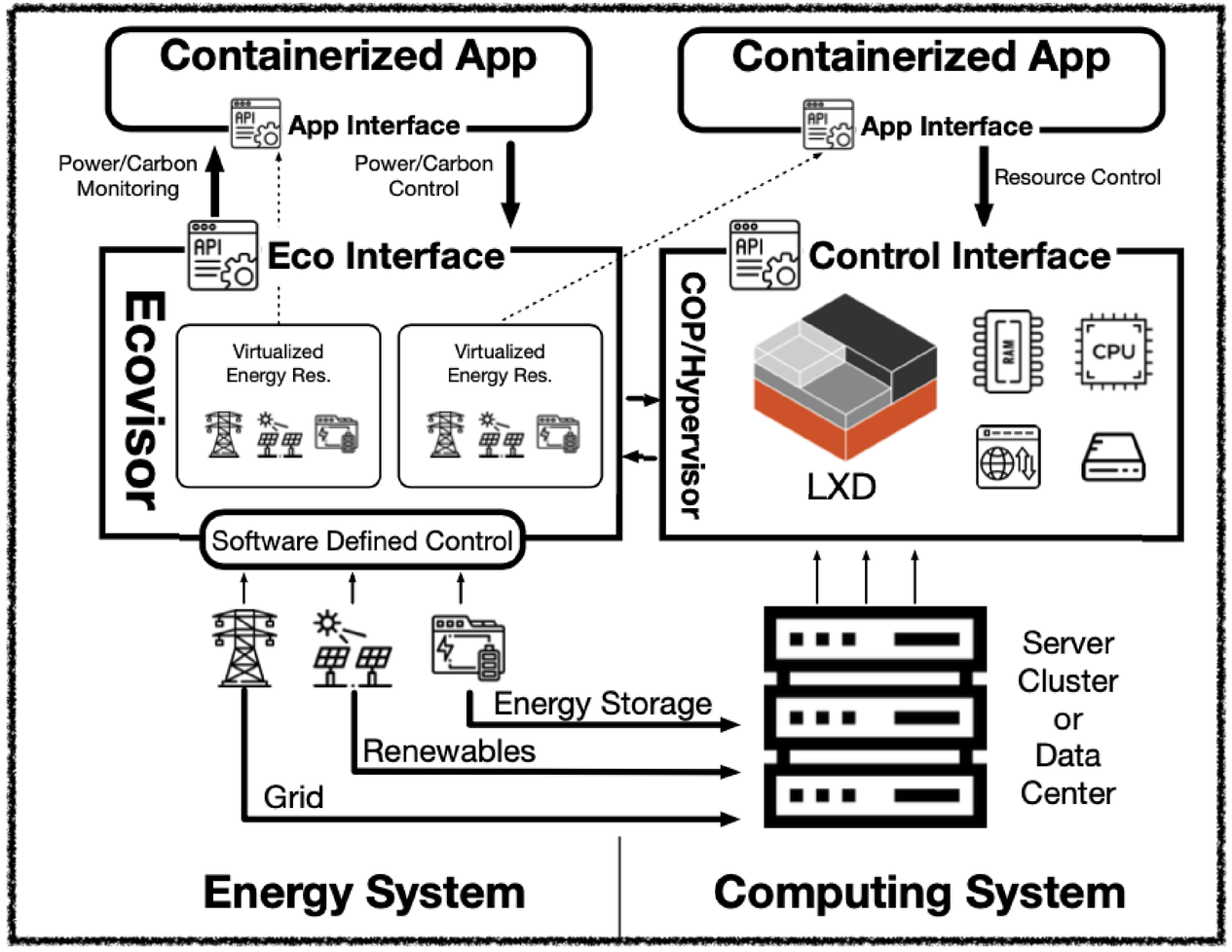
APPLICATIO
N

APPLICATIO
N

Ecovisor: Visibility and Control of Grid Dynamics



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

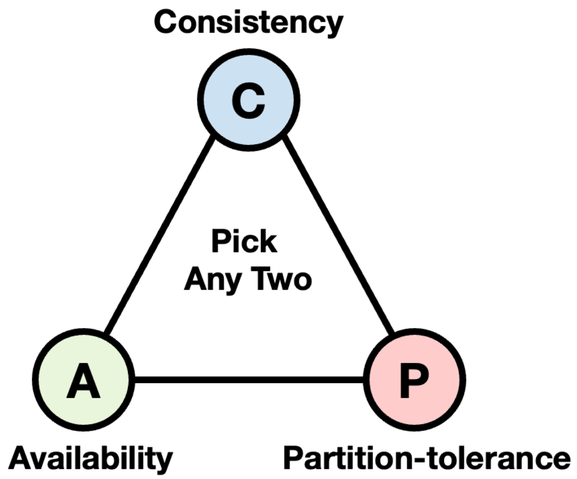
Control Power Supply and Demand

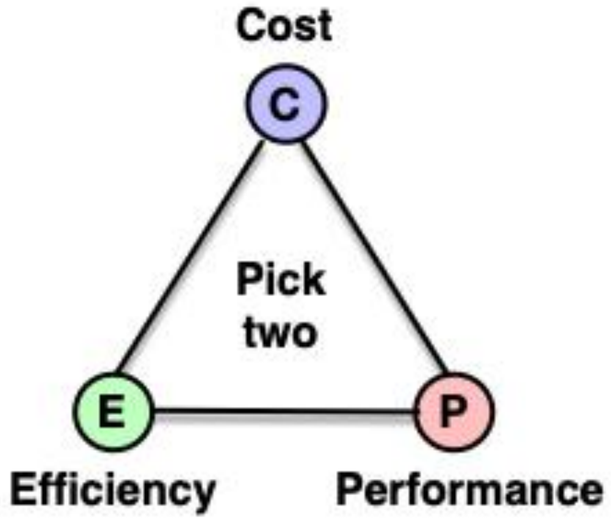
Asynchronous Notifications

Get Energy System Information

No Free Lunch: Cost, Efficiency, and Performance Tradeoffs

CAP Theorem

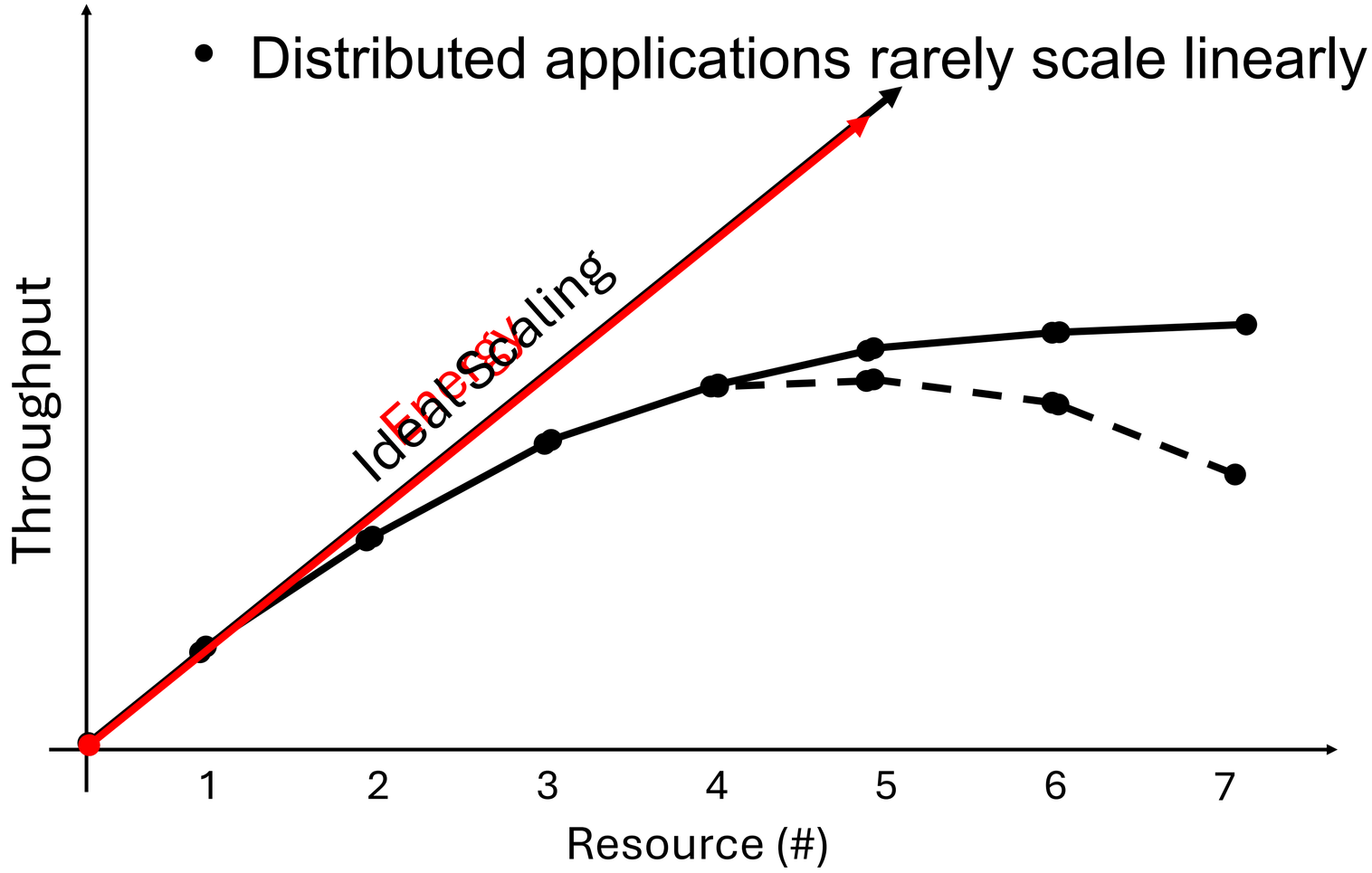




- Grid-friendly optimizations do not come for “free”
- They can decrease performance or increase cost
- **CEP Conjecture**: It may not be possible to simultaneously optimize cost, efficiency and performance in grid-friendly workload shifting

Energy-Cost Tradeoffs In Resource Scaling

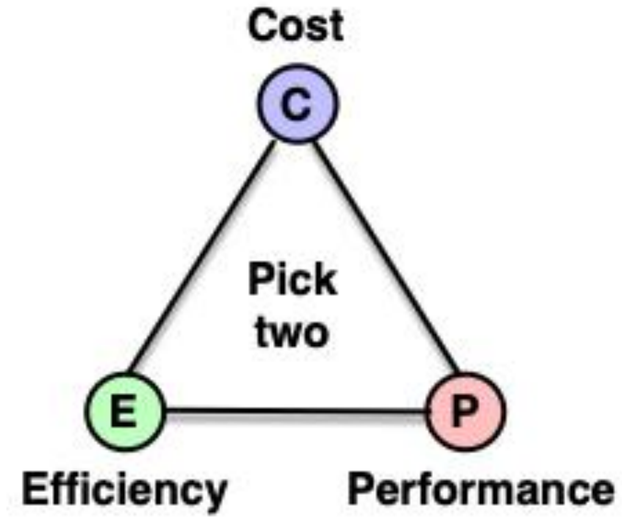
- Distributed applications rarely scale linearly



- Dimi ing throughput.
  - Efficiency
  - Higher Scale
  - ⇓
  - High Compute Time (Cost)
- Lo throughput per Joule.
  - Higher Scale
  - ⇓
  - Cost
  - Higher Energy Cons tion

CEP Tradeoffs - Rich Space for Systems Research

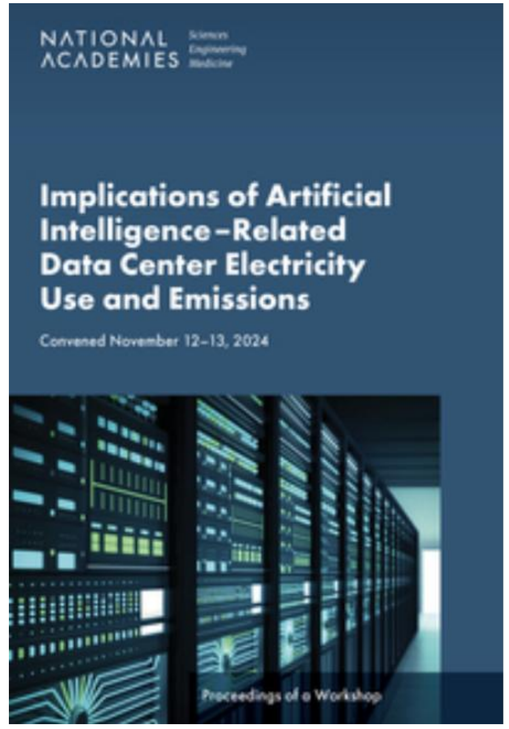
- System design under CEP tradeoffs
- Application-specific approaches
- Cloud versus edge grid-friendly approaches



National Academy Efforts

- Oct ’24: Workshop on AI, Data center, Electricity Use

- Proceedings have many good research ideas



Summary

- Future computing systems will need to be grid-friendly
  - Different from traditional energy management
- Adaptation is key - workload flexibility can help
- Rich design space for systems research with many open challenges

Thank you

EXPEDITIONS:
CoDec

- https://codecexp.us