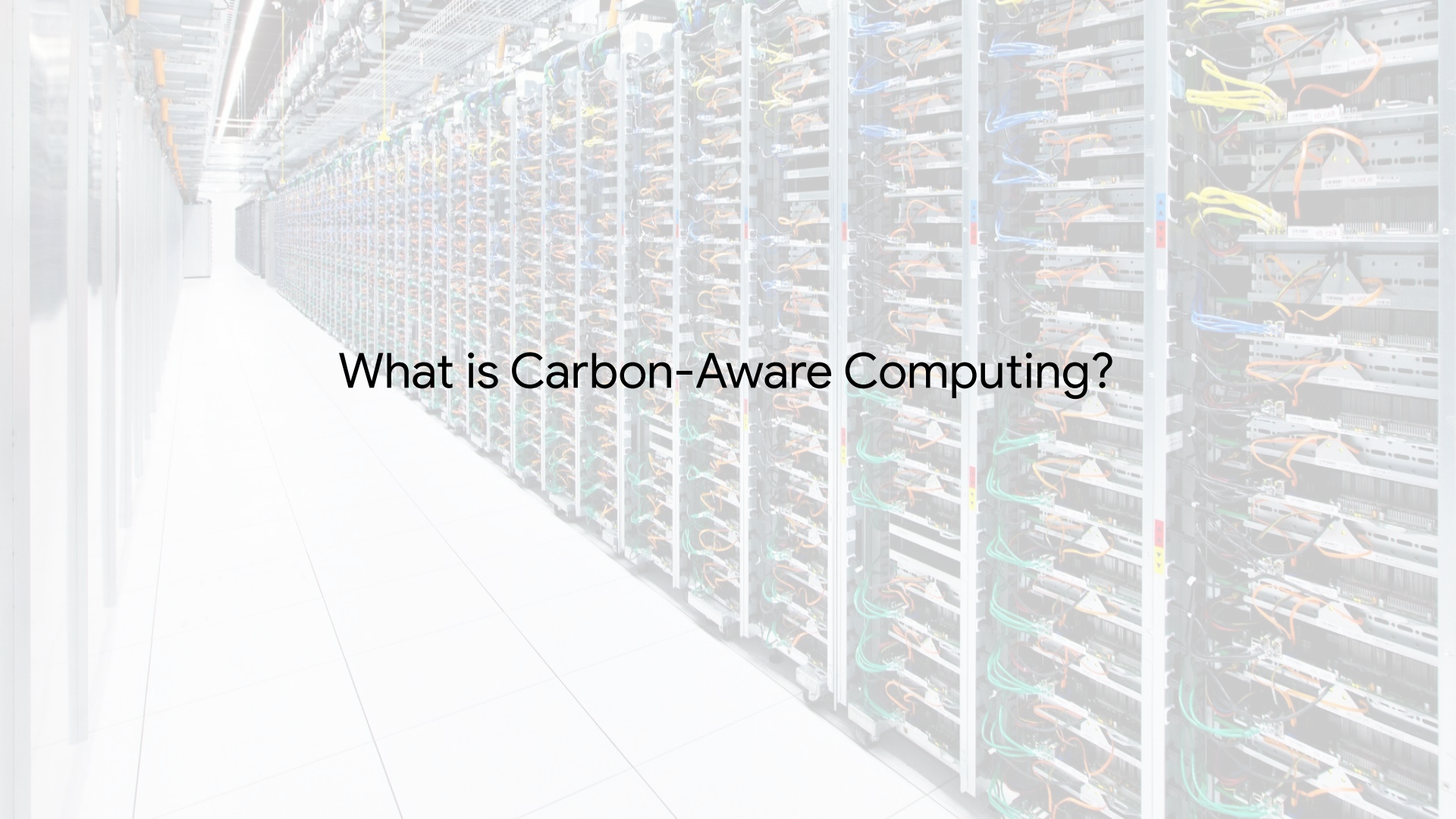
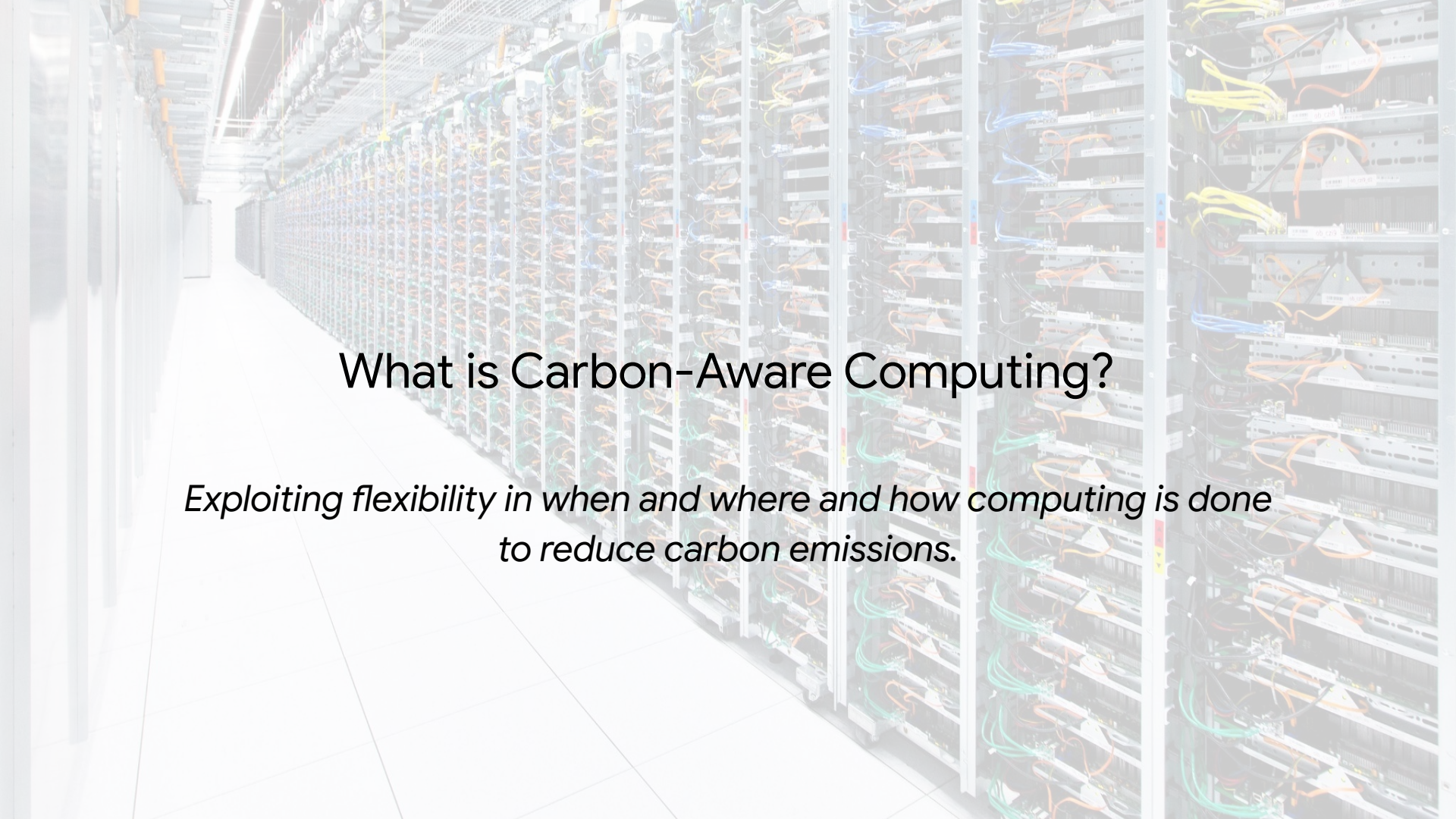
Carbon Aware Computing at Google, and Beyond

Ana Radovanovic,
Technical Lead for Carbon Aware Computing @ Google

IFIP Performance 2021



What is Carbon-Aware Computing?

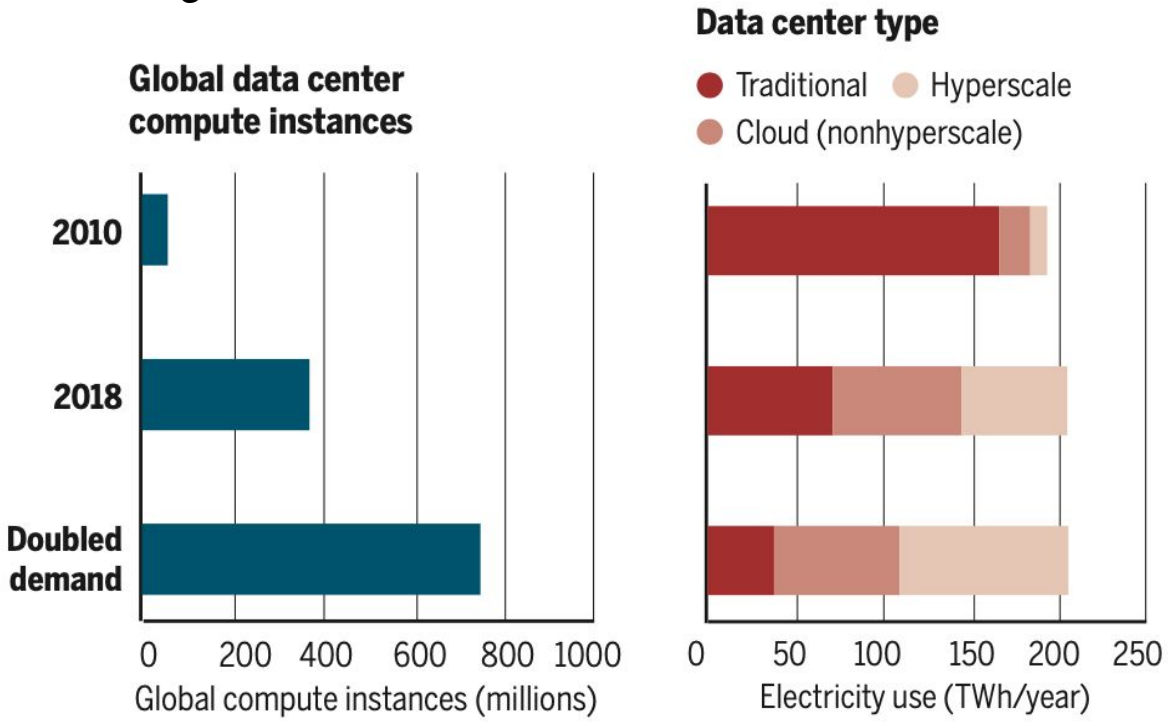


What is Carbon-Aware Computing?

Exploiting flexibility in when and where and how computing is done to reduce carbon emissions.

Why shifting datacenter loads can matter

"Demand for computing resources and datacenter power worldwide has been continuously growing to approximately 1% of the total electricity usage...in the short timespan between 2010 and 2018, global datacenter workloads and compute instances increased more than sixfold, and continue to grow..."



Cloud and hyperscale computing datacenters with some load flexibility can change their load pattern on the grid. This can influence electricity markets and hence carbon intensity.

***Recalibrating global data center energy-use estimates**, Masanet et. al., Science 2020.*

Some computing has flexibility in when it can run ...

Examples include data compaction and distributed computation for

- Processing videos
- Feature extraction and training large-scale machine learning models to optimize Web search, content & social recommendation systems, etc.
- Simulation pipelines
- ...and many other **latency-tolerant** workloads (e.g., batch)

… so load profile can be shaped over a 24 hour period

Some computing jobs have flexibility where they can run …

Examples include: user-facing services that can be **geographically rebalanced** or **resourced**

- *Applications*: Microsoft's Office, Facebook & Twitter messaging, Salesforce's enterprise solutions, Google's G Suite, YouTube, Maps, etc.
- *Compute resources*: AWS, Azure, G Cloud, IBM Cloud, …
  - Carbon-aware:
    - Microsoft's Carbon-Aware Kubernetes
    - Google Cloud's Region Picker

…and we hope to see more

… so load profile can be shaped across data centers

The idea, in theory, is not entirely new ...

Theoretical treatments, small-scale prototypes and simulation-based studies

- *Compute shifting across space*
  - Le et. al. [2010], Liu et. al. [2011/2014], Ren et. al. [2012], Berral et.al. [2014], Rahman et. al. [2014], Deng et. al. [2016], Kelly et. al. [2016], James et. al. [2019], Zheng et. al. [2020], ...
- *Compute shifting in time*
  - Goiri et. al. [2011/2012], Liu et. al. [2012], ...
- *Datacenter-grid integration (grid coupling)*
  - Lin et. al. [2021]

New theory, data, and engineering to realize CICS

The flexibility is enabled by

- Accurate carbon intensity data (Tomorrow)
- Scalable infrastructures (Cloud)
- Virtualizations and migration mechanisms (VMs)
- Well identified flexible load
- Global optimization using new, data-driven methodology
    - "**Carbon-Aware Computing for Datacenters**", submitted to IEEE Transactions on Power Systems, 2021.

The first deployed systems

- Google's ***Carbon-Intelligent Computing Platform*** [April 2020]
- Microsoft's ***Carbon Aware Kubernetes*** [October 2020]
- Google's ***Carbon-Intelligent Computing Platform*** shifts compute across datacenters [May 2021]

Vision: Use load flexibility to reduce emissions

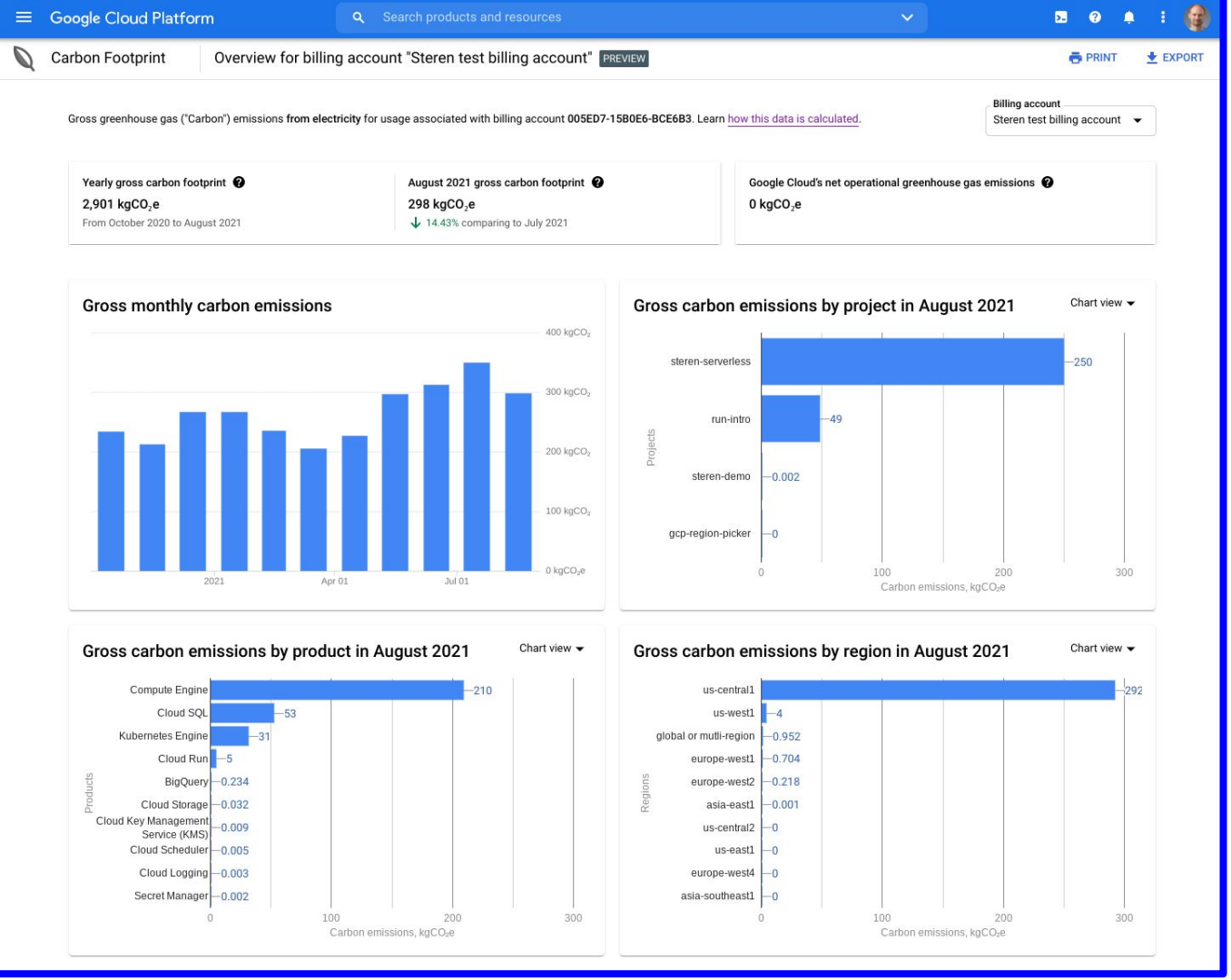
Pillars of carbon-awareness

User facing: Carbon reporting

Developer facing: Carbon signaling

Infrastructure: Flexible load management

(Re)Engineering: Increasing amount of flexible load



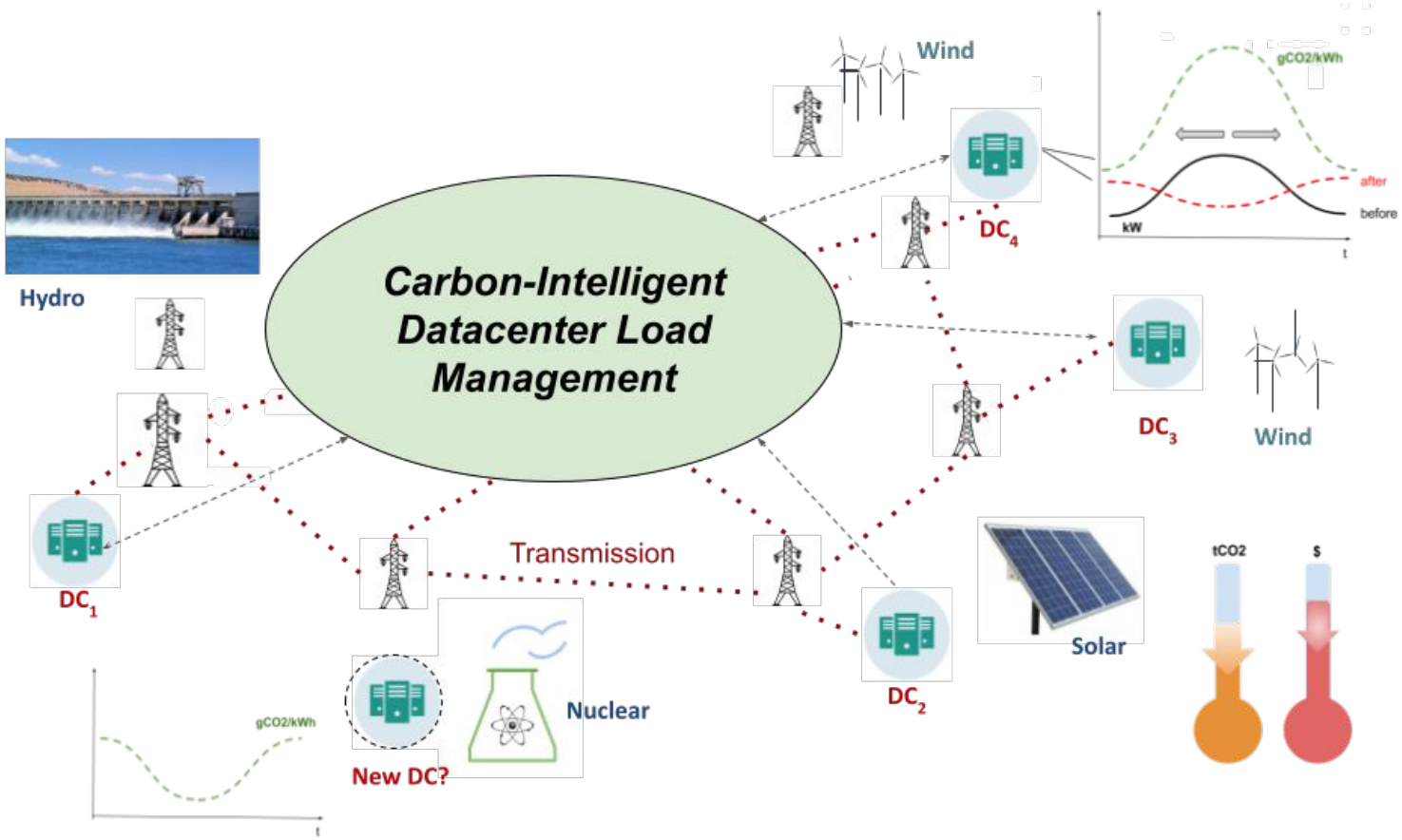
Carbon reporting

Google Cloud customers can choose “greener” cloud locations

Google’s approach to flexible load management

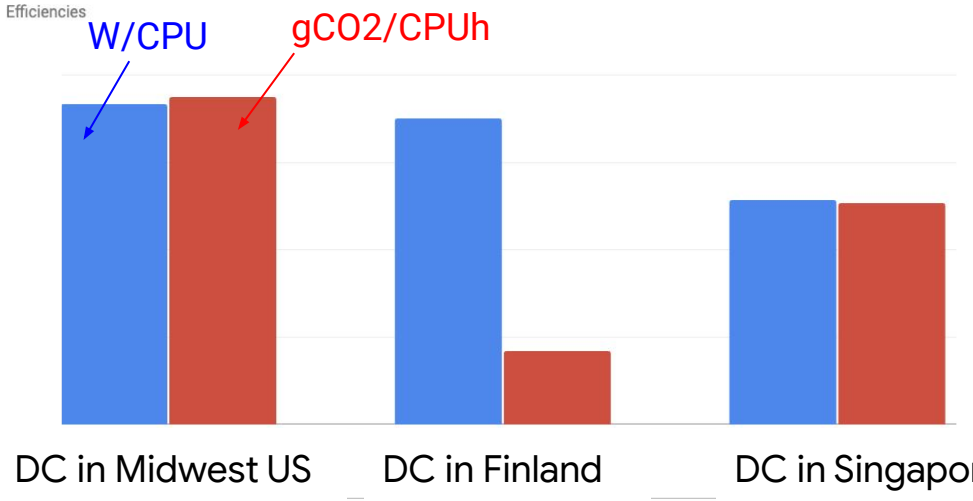
Flexible load management

Intelligent management entails that both workload flexibility and electricity markets dynamics need to be harnessed to reduce net total CO_2 emissions.

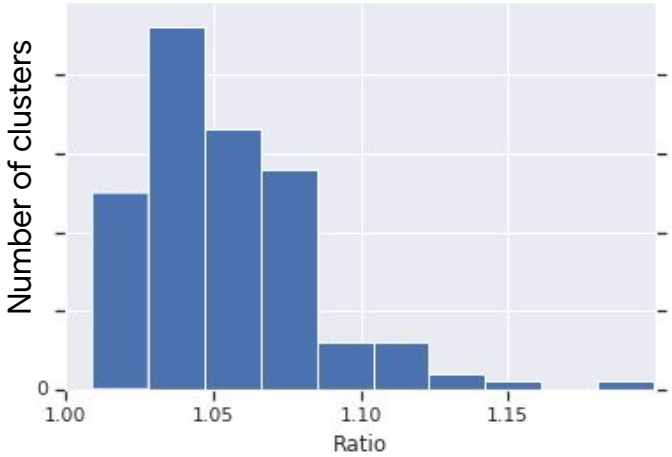


Real-world opportunities

Datacenter hardware can significantly affect power usage per unit of compute work done

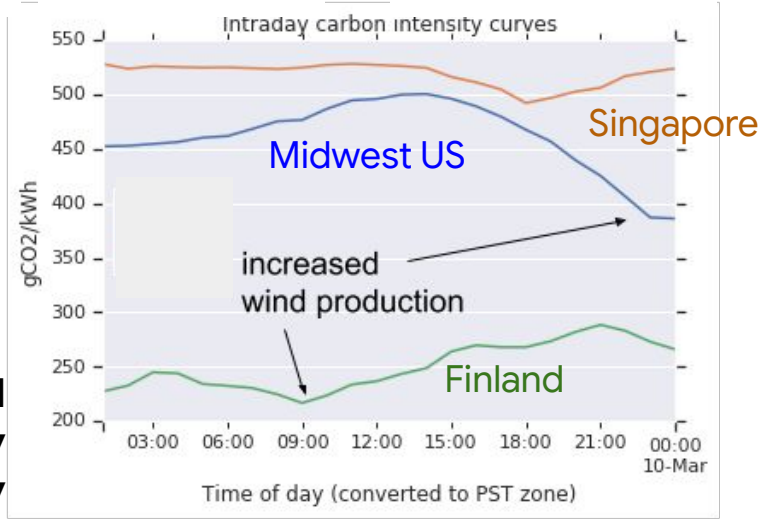


Histogram of cluster level peak to average power ratios



Daily peak-to-mean power usage ratio differs across clusters

Local grid-level CO_2 *varies* by time of day



Google's load shaping: challenges and a key opportunity

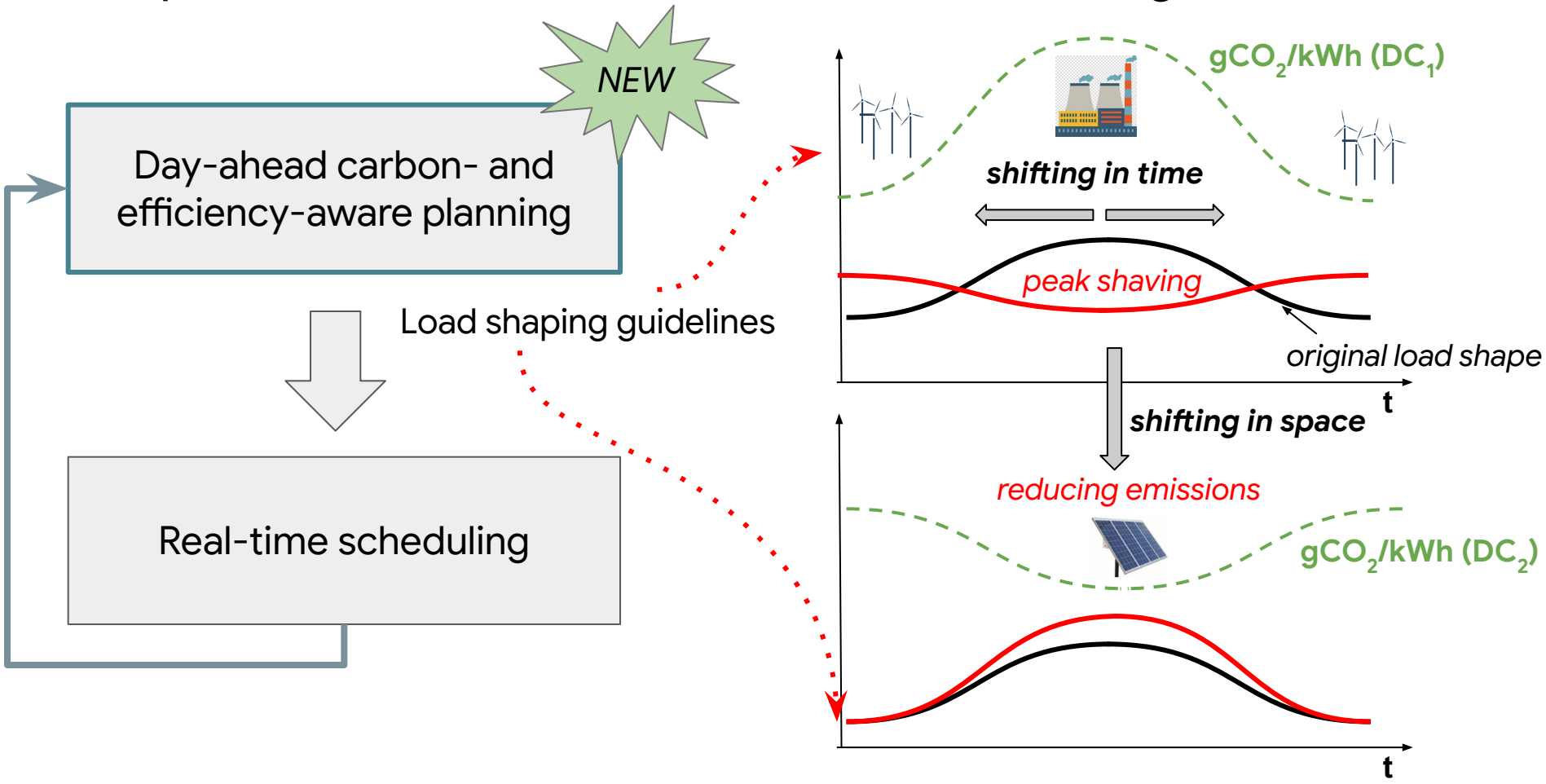
Challenges for moving load in space and time

- Load management must meet Google's reliability principles
- Hard infrastructure constraints (machine capacity, circuit breaker limits, etc.)
- Workload performance expectations need to be met
- Job placements have dependencies and consequences
- Compute jobs' resource demands have uncertainties
- Real-time job scheduler complexity has to be as low as possible

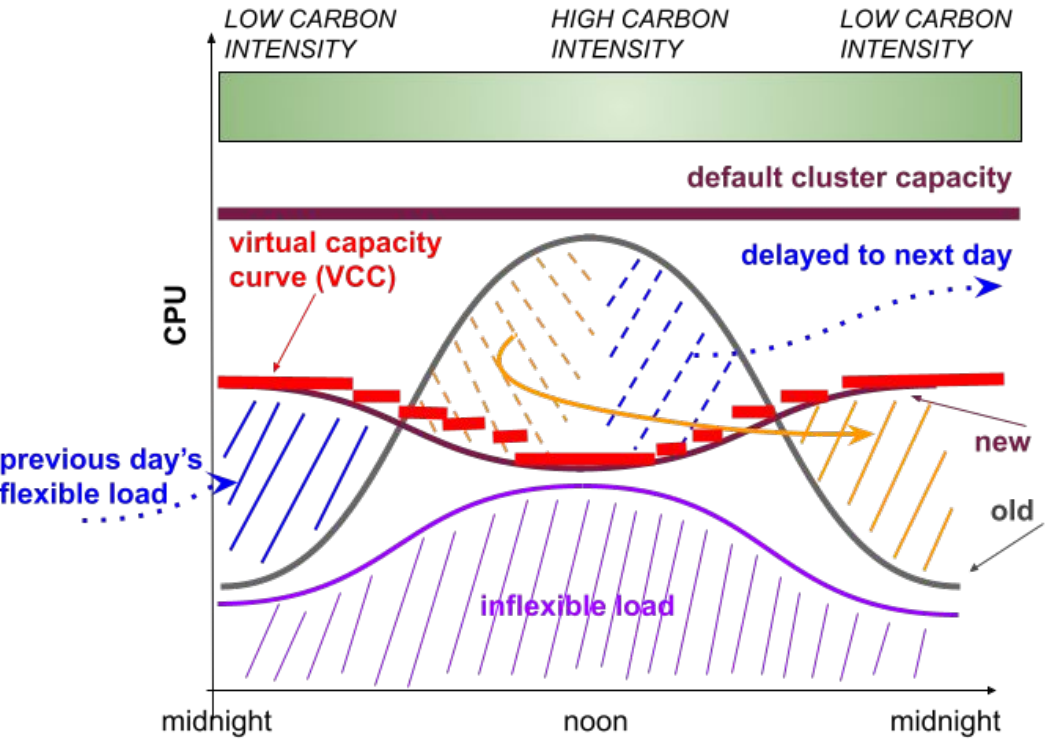
But what fundamentally makes our approach tractable is that

- The total amount of work that needs to get done per day is quite predictable
  - Consequence: total load envelope can have different shapes, so let's pick an optimal shape

Conceptual division for effective workload management



Load shaping: applying Virtual Capacity at the cluster level



A 24-hour ***Virtual Capacity Curve (VCC)*** shapes each cluster's load

- The real-time scheduler will postpone or move flexible workloads that would cause the virtual capacity to be exceeded
- Inflexible workload is not affected

The combined impact of these VCCs makes Google's workload both carbon-aware and resource efficient.

What makes this approach effective at Google?

Day-ahead demand and grid-level carbon intensities are predictable

- Intraday inflexible as well as daily flexible compute usage
- Compute usage -> power usage
- Average carbon intensities (ElectricityMap.org)

Scalability: Load shaping acts as a simple constraint to real-time scheduling

- Scalable optimization framework that uses aggregate demand estimates to produce capacity curves (shaping guidelines) for all clusters fleetwide

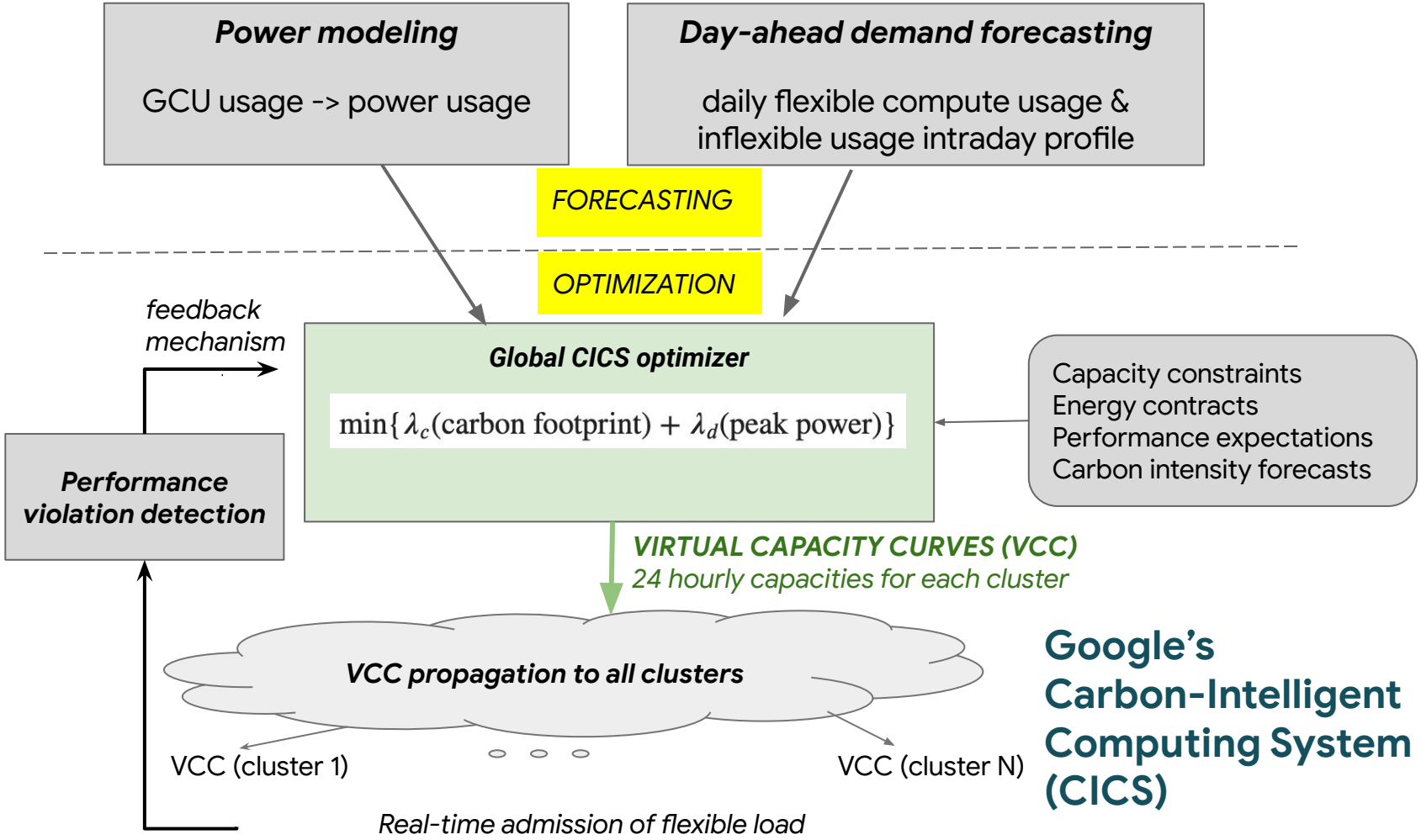
Computationally efficient, scalable & extensible to future use cases

- e.g., enhanced spatial shifting, portfolio level optimizations, etc.

It is possible to incorporate risk-awareness

VCC computation incorporates

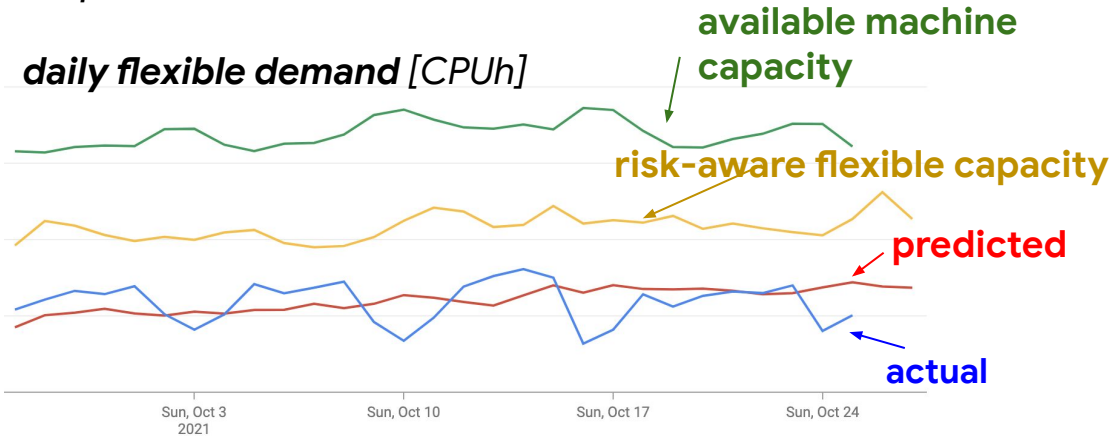
- Explicit workload performance expectations
- Prediction uncertainty of resource/power demand
- Power infrastructure limits, limits set by energy contracts, etc.



Google's Carbon-Intelligent Computing System (CICS)

Daily forecasts for a cluster

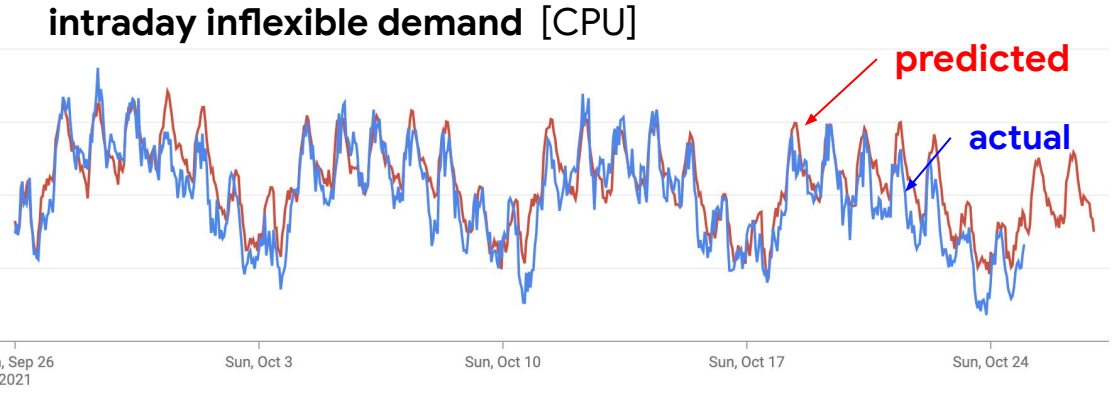
How much load flexibility can be expected in a given cluster? What will be the inflexible load shape?



Median APE*:

- **< 20% for more than 90%,**
- **< 10% for ~60%**

of Google clusters globally.

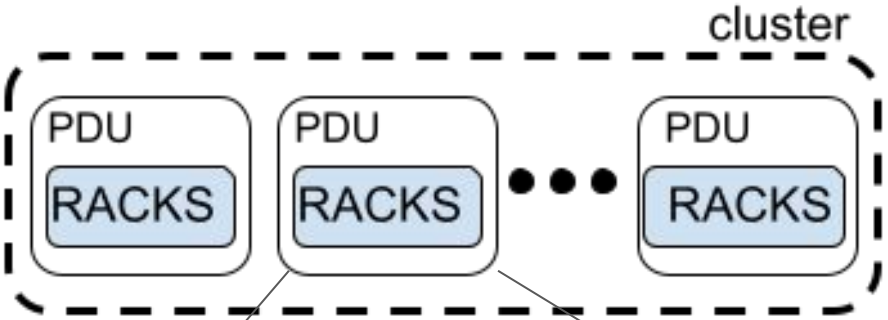


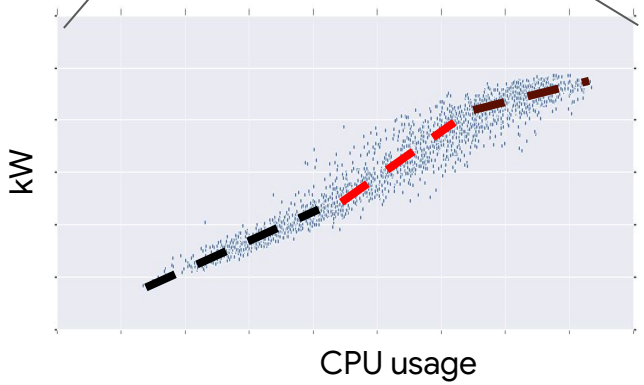
Median APE is smaller than 8% for more than 90% of Google clusters globally.

* APE = Absolute Percent Error

Power models

How does change in CPU usage translates into change in kW?

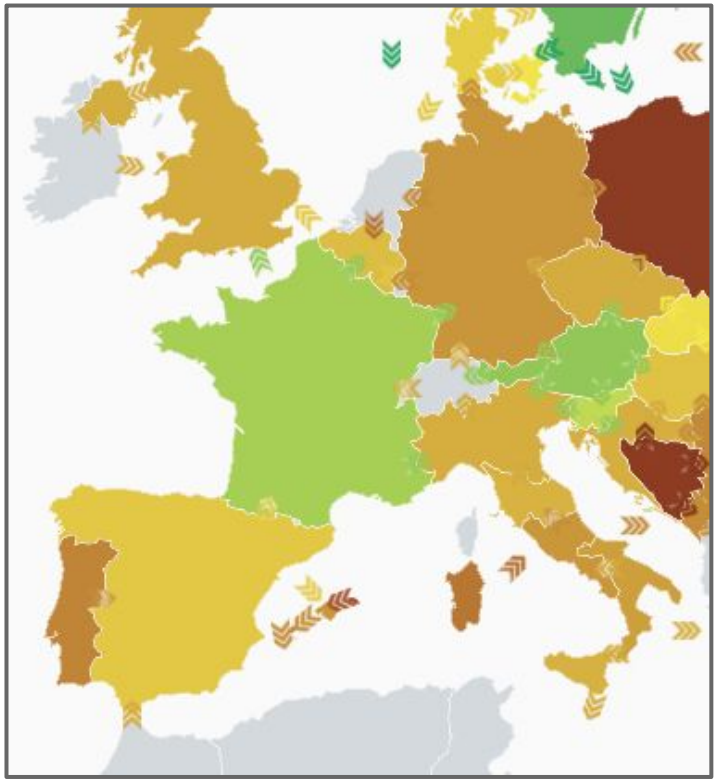




Mean APE is smaller than 5% for more than 95% of all Google's power domains (PDUs) globally.

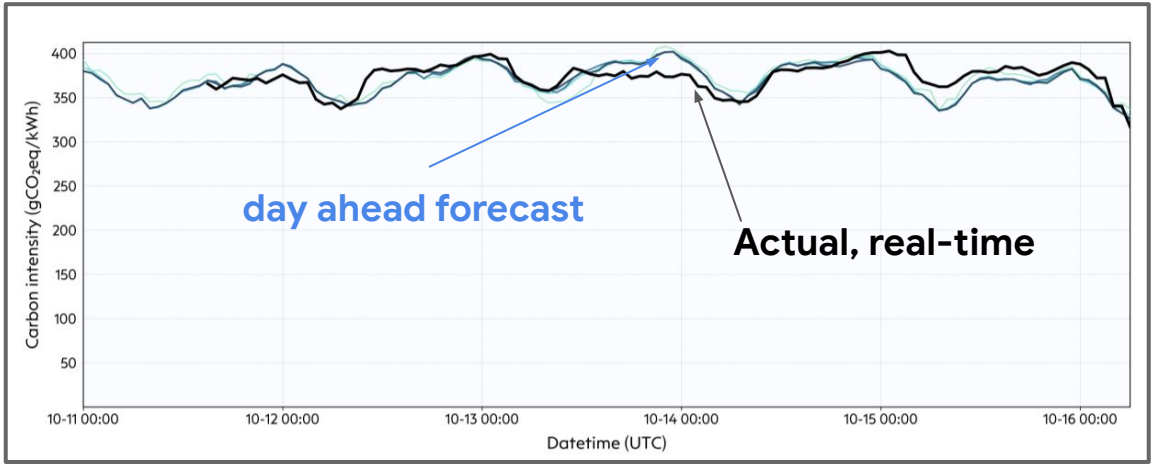
Power domain power can be accurately modeled using piecewise linear models.

ElectricityMap API: Day-Ahead Carbon-Intensity



Tomorrow Inc. provides an online map view of electric grid carbon intensity

An API provides access to an hourly day-ahead prediction of the carbon intensity for each grid region.



The math: convex co-optimization

$$\min\{\lambda_e(\text{carbon footprint}) + \lambda_p(\text{peak power})\}$$

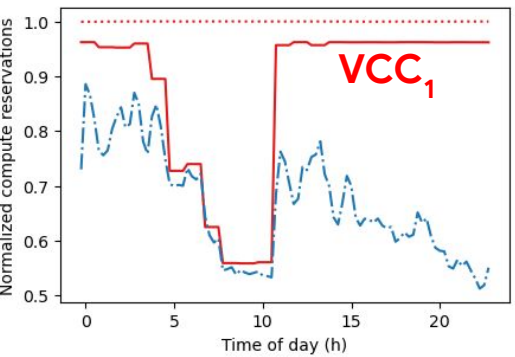
Co-optimization of **CO_2 impact** and **Infrastructure cost**

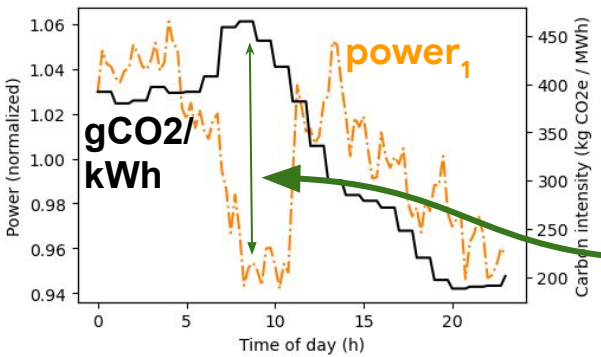
- We can **select the trade-off** between the two: like an internal carbon price
- Much of the time, these don't fight each other - we see many examples where **both good behaviors are observed**
- This central optimization performed daily has **fleetwide impact**

Demonstrations of real life impact

Impact of CICS shaping depends on

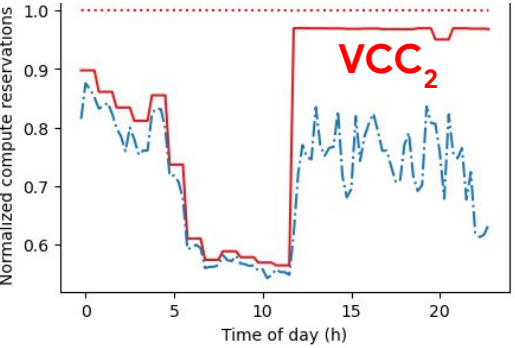
- amount of flexible load
- prediction uncertainty
- variability and magnitude of carbon intensity forecasts

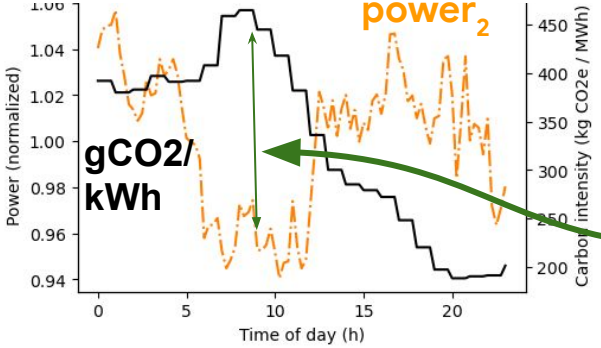




High uncertainty, large flexible usage

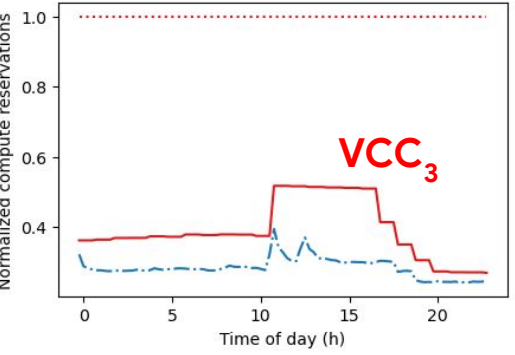
~8% power usage drop for **3 hours** at peak carbon intensity hours

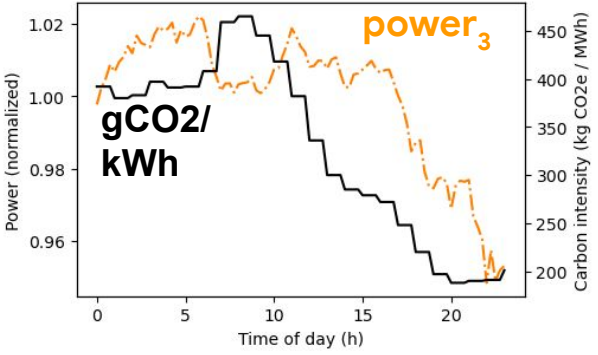




Moderate uncertainty, large flexible usage

~8% power usage drop for **6 hours** at peak carbon intensity hours

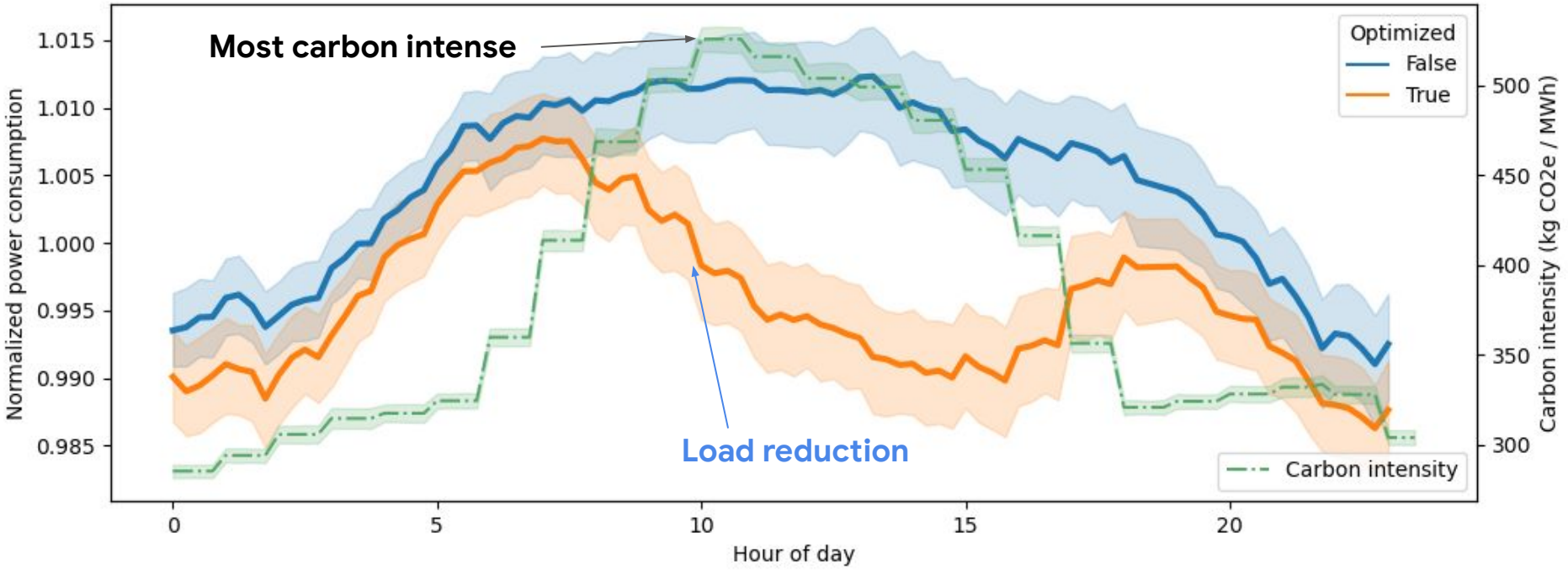




High uncertainty, small flexible usage

No meaningful curtailment of power usage during peak carbon intensity hours

Example of a datacenter exposed to CICS shaping



The effect of load shaping optimization on cluster power consumption, showing reduction in power consumption during period of higher carbon intensity

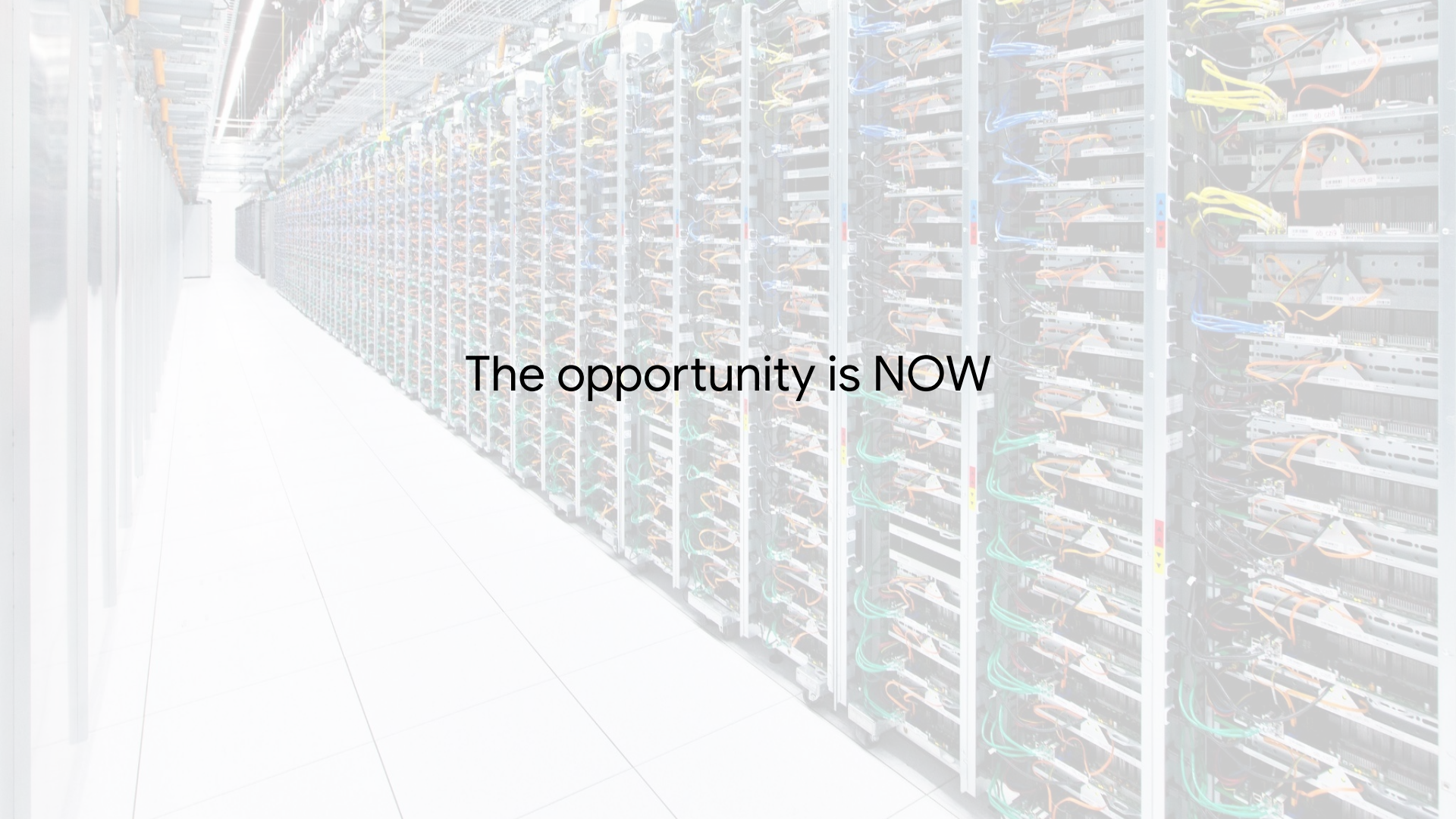
Open challenges

Understanding flexibility

Effect of load shaping on workloads

Ensuring that spatially flexible load ends up at the “right” location

How does load shaping affect CO_2 emissions?


The opportunity is NOW

Some opportunities

Embed carbon signals into cloud products

Steer web (e.g. search) requests to “greener” locations

Build tools to identify flexible compute workloads

(Re-)Engineer software so that parts are more flexible in time and space

Migrate applications to “greener” cloud regions

Carbon-aware cloud-controlled devices (not only compute)