Scheduling and Placement for Low-Carbon Edge/Cloud Computing

March 17, 2022

LCSC Seminar Series

Dr. Lauritz Thamsen

UofG – SoCS – GLASS

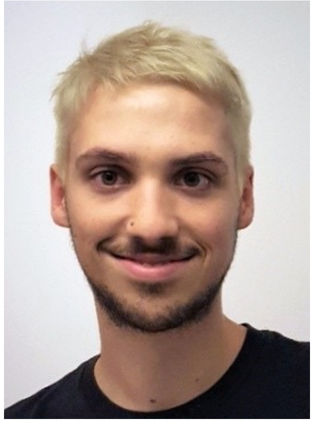
https://lauritzthamsen.org

Outline

- Background
- Simulation of Energy Consumption
- Temporal Cloud Workload Shifting
- Forecast-Based Admission Control
- Outlook & Summary

Acknowledgement

- Work done with PhD students in Prof. Kao's Distributed and Operating Systems group at TU Berlin



Philipp
Wiesner

Dominik
Scheinert

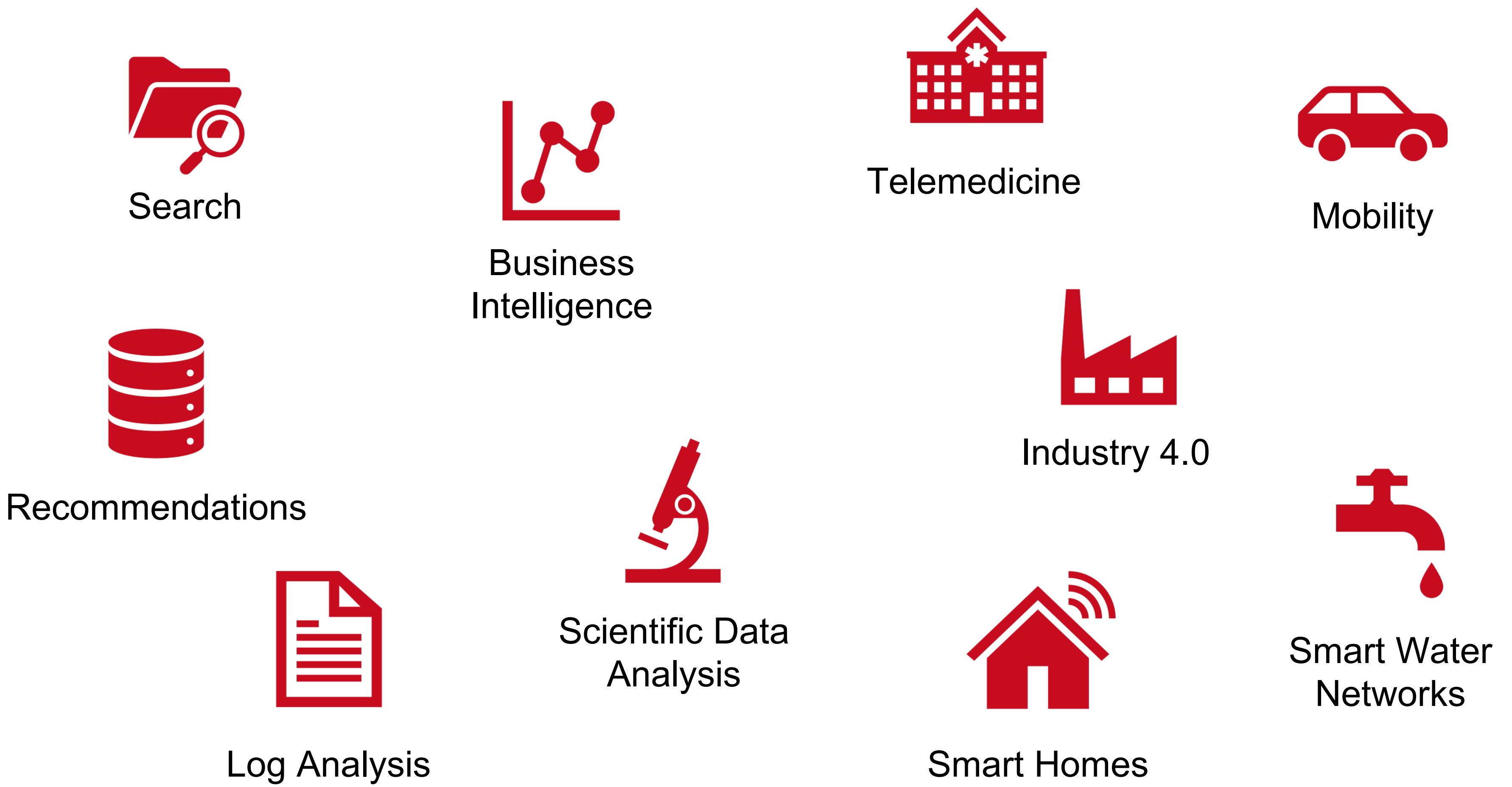
Kordian
Gontarska

Ilja
Behnke

Thorsten
Wittkopp

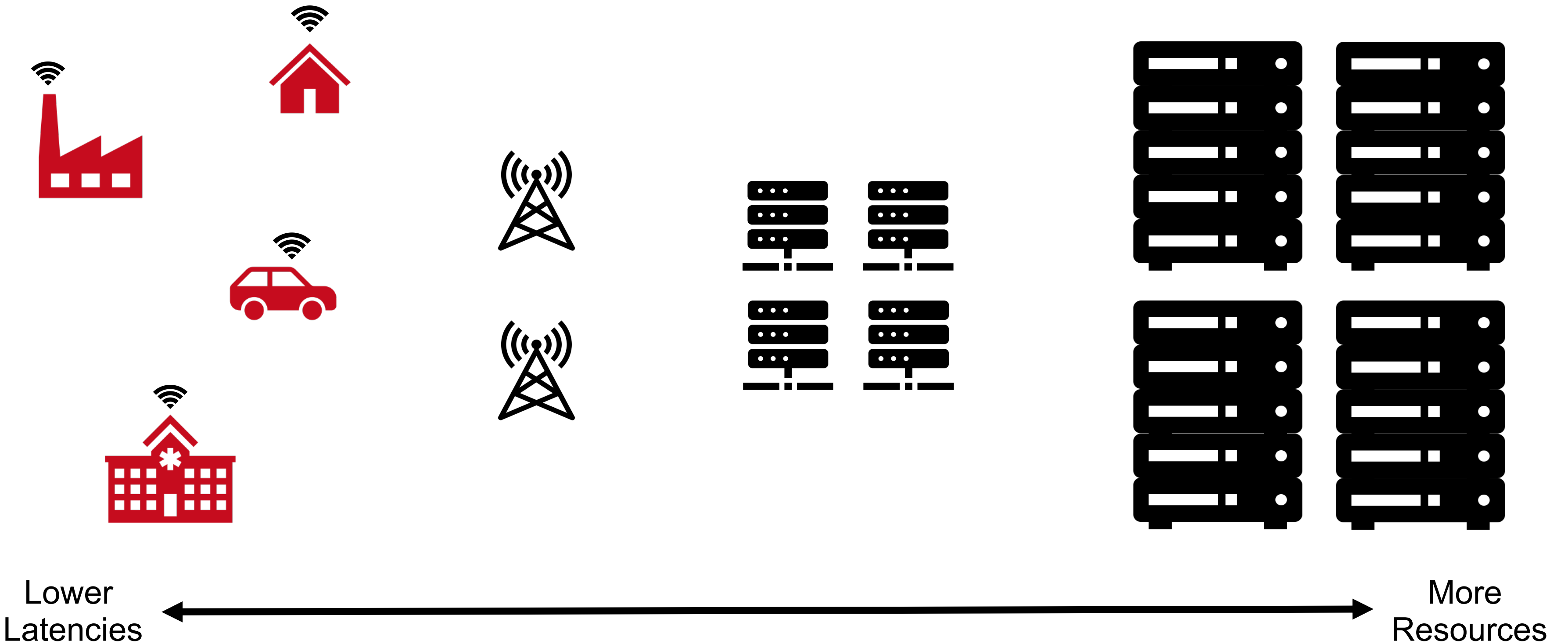
Background

Data-Intensive Applications



Diverse Computing Infrastructures

Heterogeneous and dynamic distributed computing environments from devices to data centers

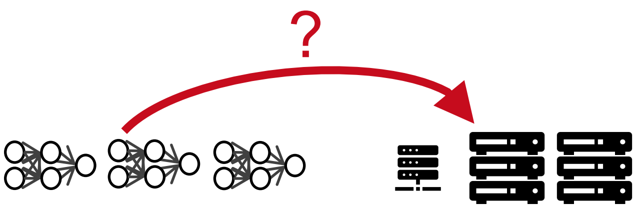


Research Questions

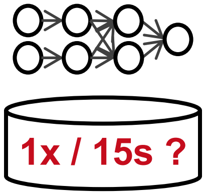
Given a job and objectives/constraints for its execution:

1. What resources to use for a job?

2. When and where to run the jobs?



3. How to set system configurations?



Sustainability Objectives

- Low energy consumption: Caching, local computing, utilizing resources fully (before scaling out more), energy-efficient languages and systems, …
- Low carbon emissions:
    - Grid energy carbon emissions are often different over time and locations → use *low-carbon energy*
    - On-site renewable energy sources → use *all* the green energy
- First requirement? Understanding energy demands…

LEAF: Energy Simulation

LEAF: Simulating Large Energy-Aware Fog Computing Environments.
Philipp Wiesner and Lauritz Thamsen. 5th IEEE International Conference on Fog and Edge Computing (ICFEC). IEEE. 2021

The "LEAF" Simulator

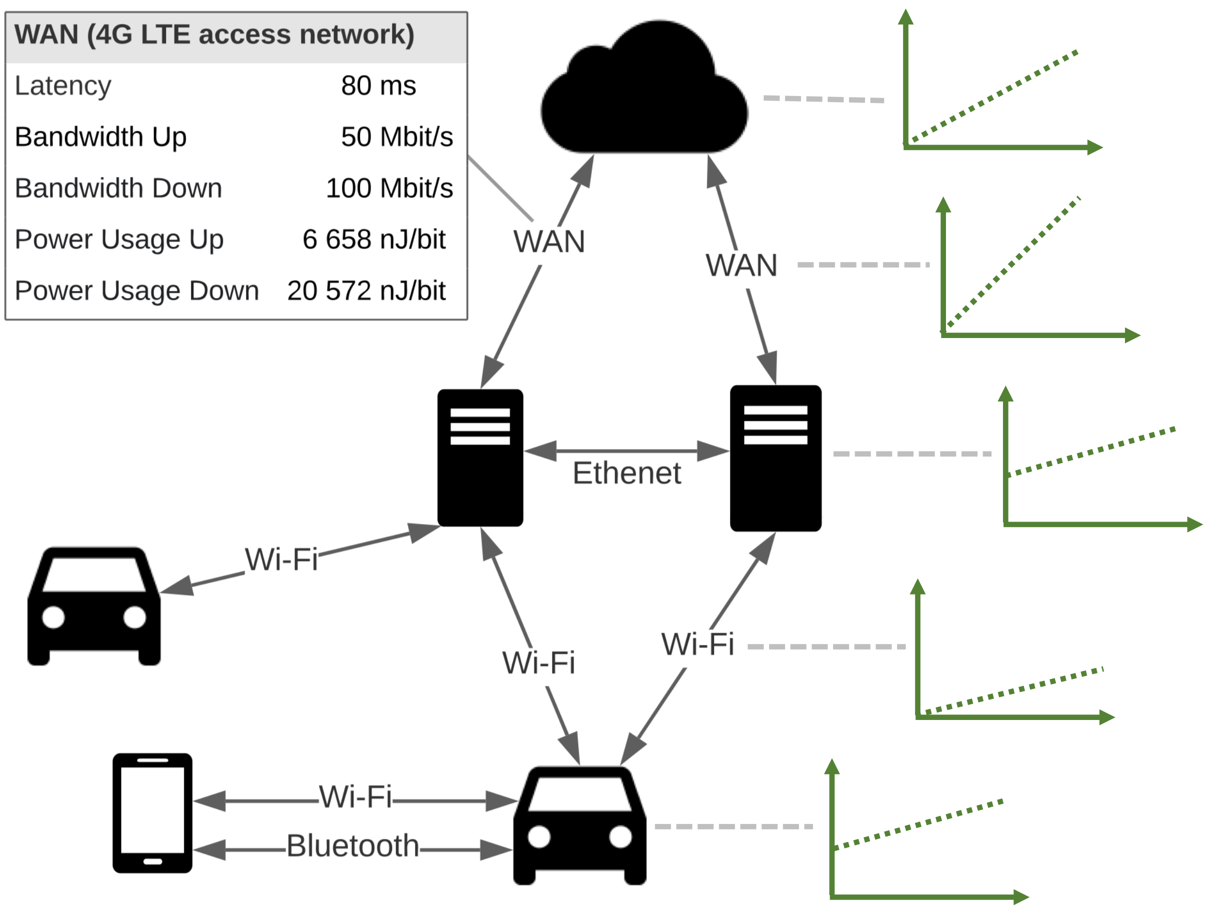
- Simulator for modeling the energy consumption of applications in cloud/fog/edge computing environments

- Design goals:
  - holistic but granular power modeling
  - realistic and dynamic compute environments
  - energy-aware online decision making
  - thousands of devices & applications

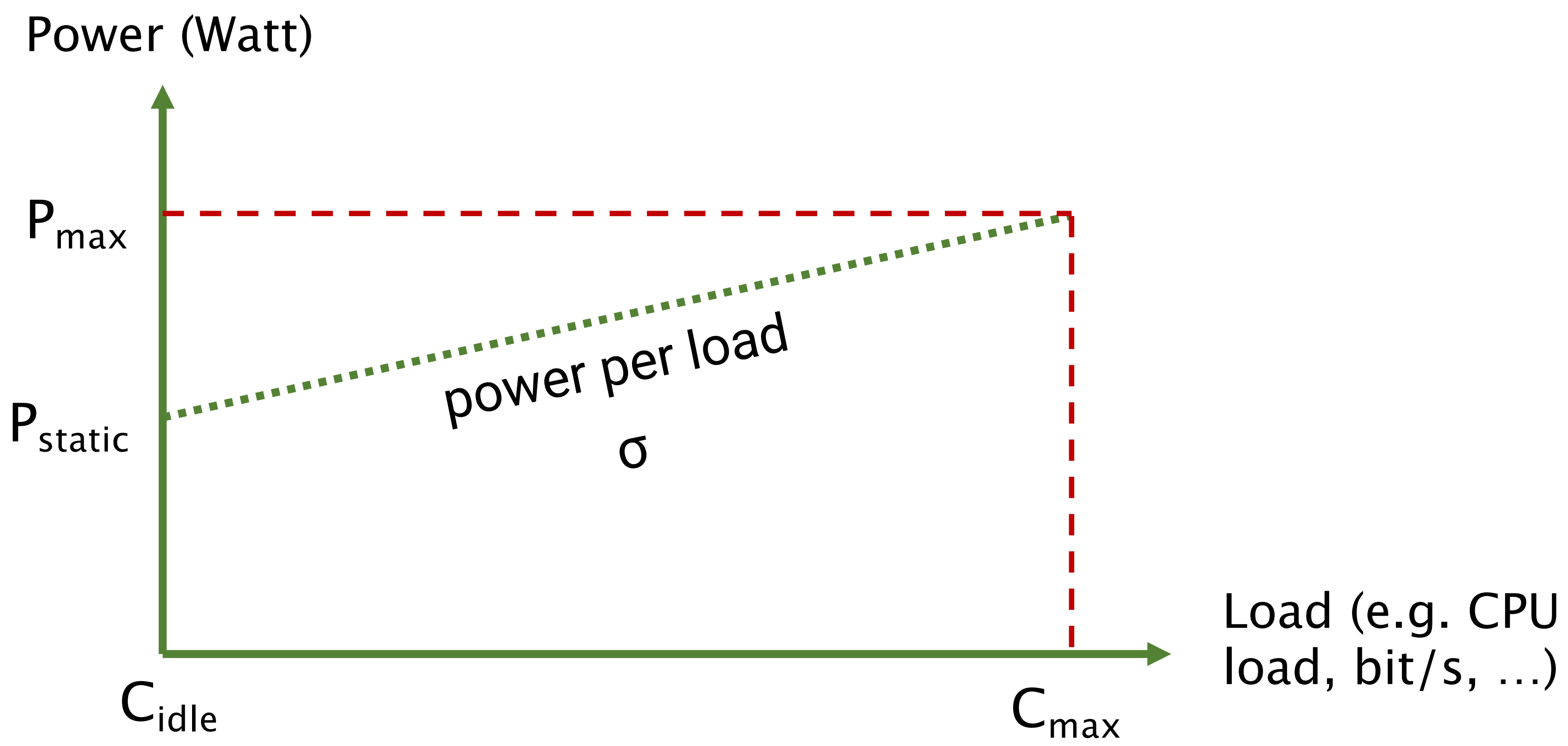
LEAF

https://github.com/dos-group/leaf

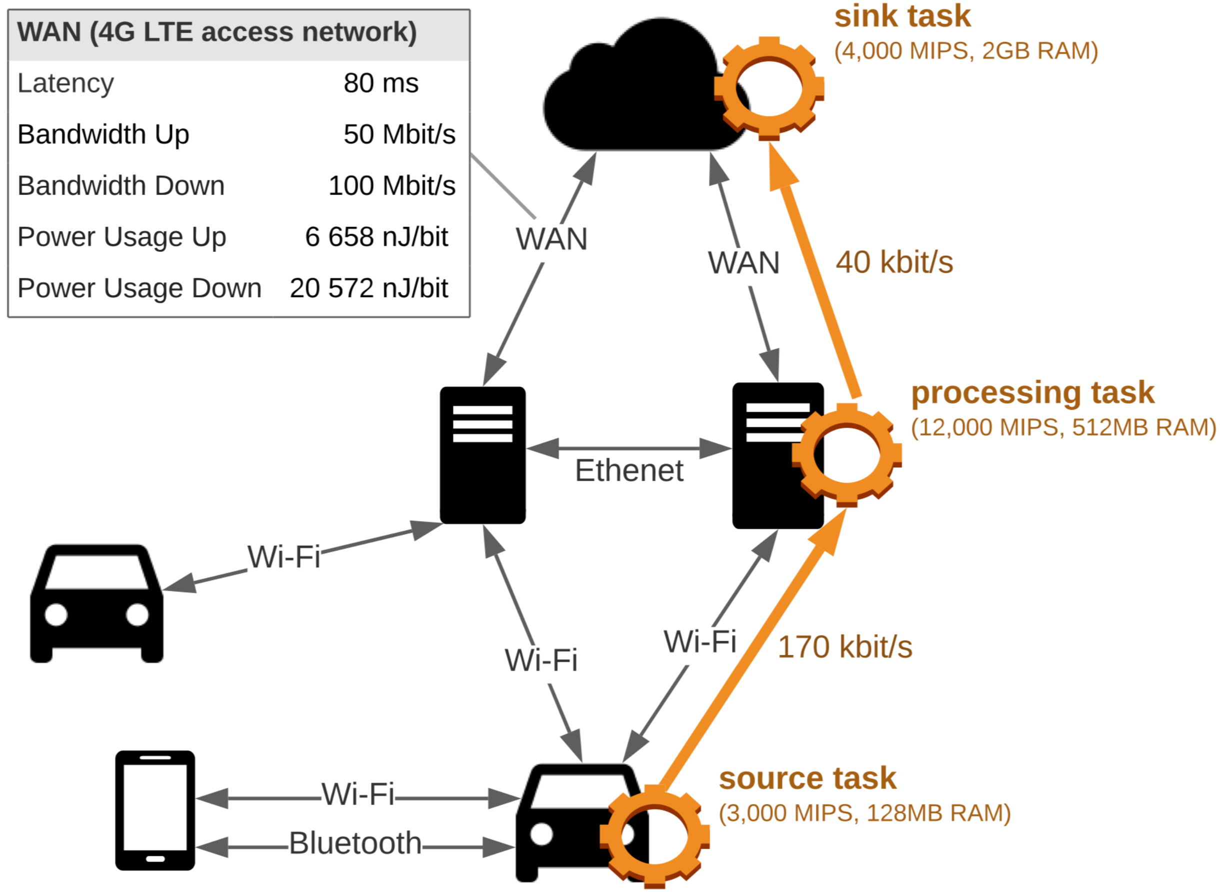
Infrastructure Model



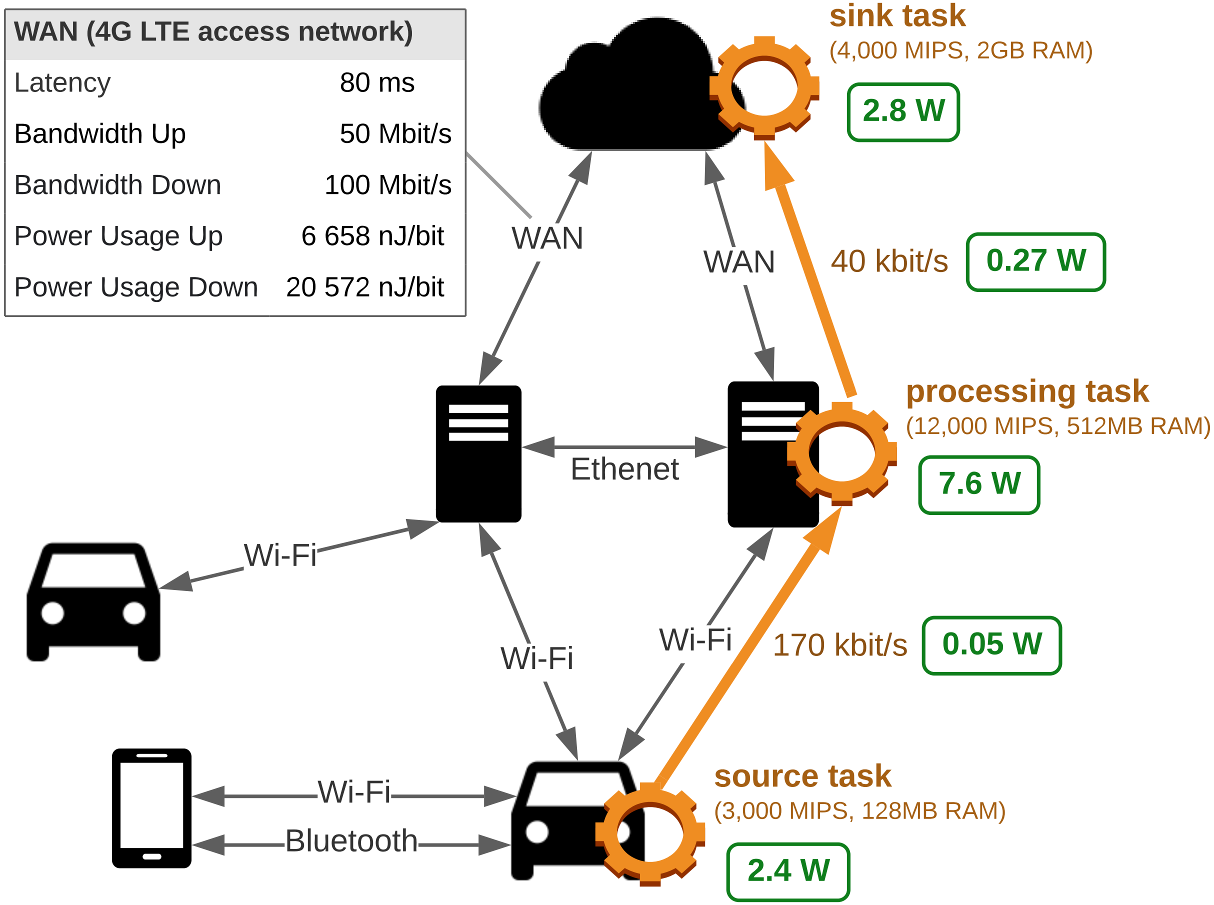
Power Modeling



Application Model

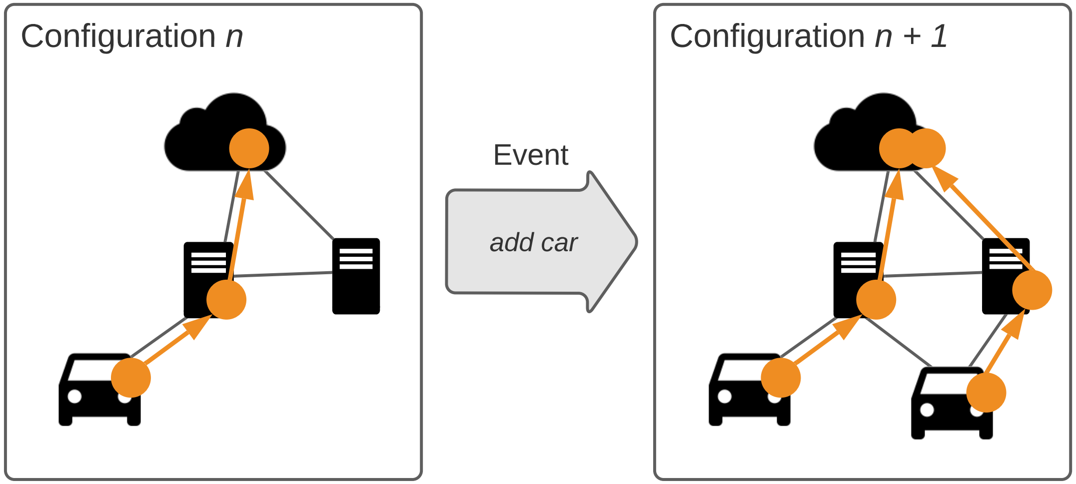


Application Power Usage



Event-Based Simulation

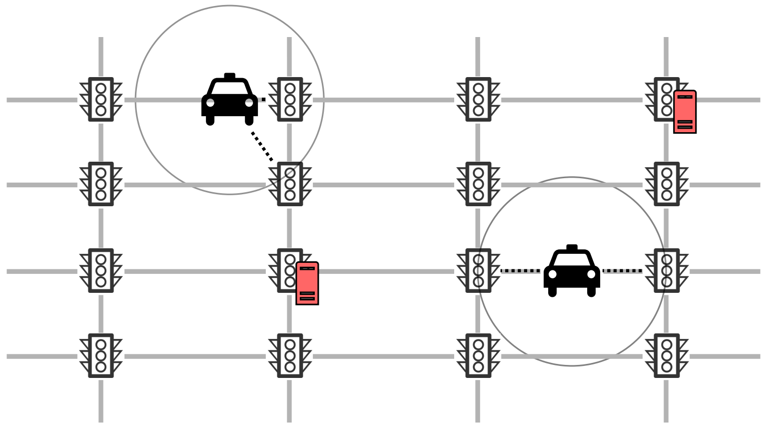
- Events read and update the infrastructure and application models

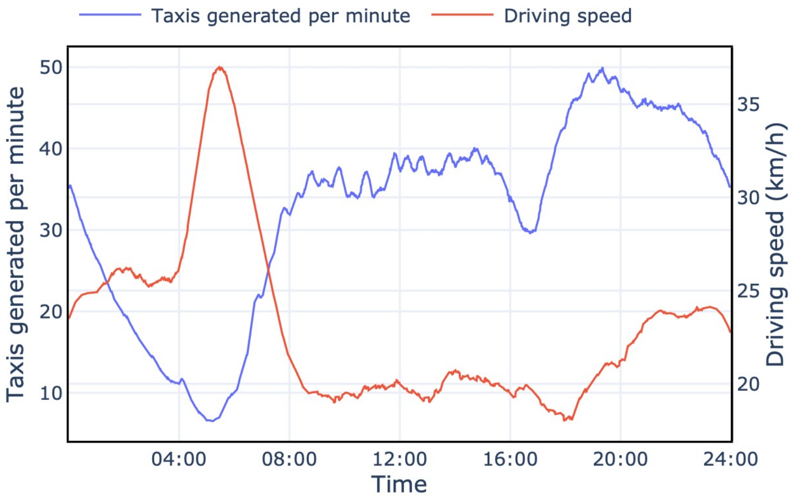


Experiment Setup

- 24 hours of taxi traffic
- time steps of one second

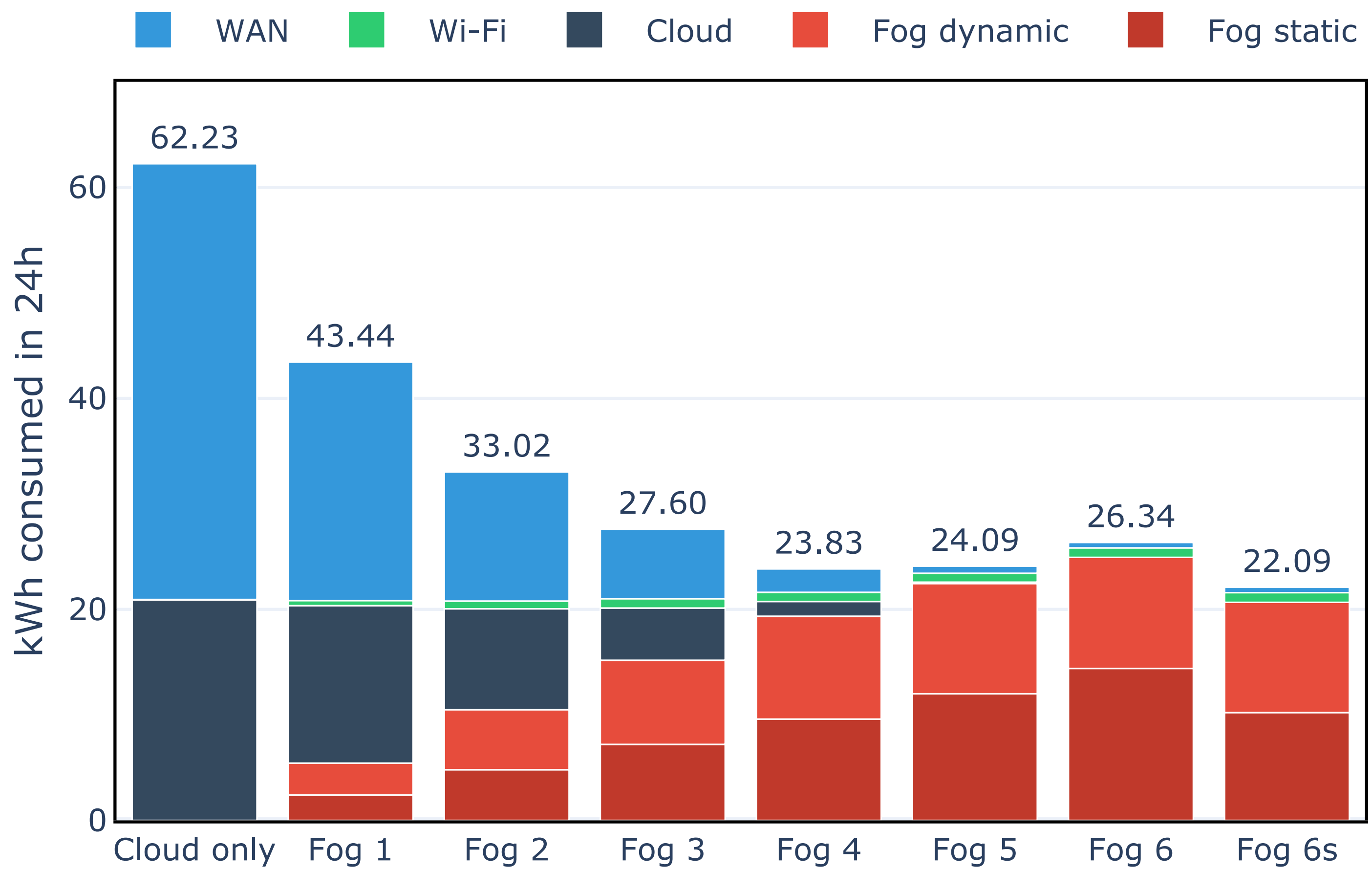
	Max load	P_{static}	σ
Cloud	∞	-	700 µW/MIPS
Fog node	400 000 MIPS	100 W	350 µW/MIPS
WAN $_{\text{STL} \rightarrow \text{Cloud}}$	50 Mbit/s	-	6658 nJ/bit
WAN $_{\text{Cloud} \rightarrow \text{STL}}$	100 Mbit/s	-	20 572 nJ/bit
Wi-Fi $_{\text{Taxi} \rightarrow \text{STL}}$	1.3 Gbit/s	-	300 nJ/bit
Wi-Fi $_{\text{STL} \rightarrow \text{STL}}$	1.3 Gbit/s	-	100 nJ/bit



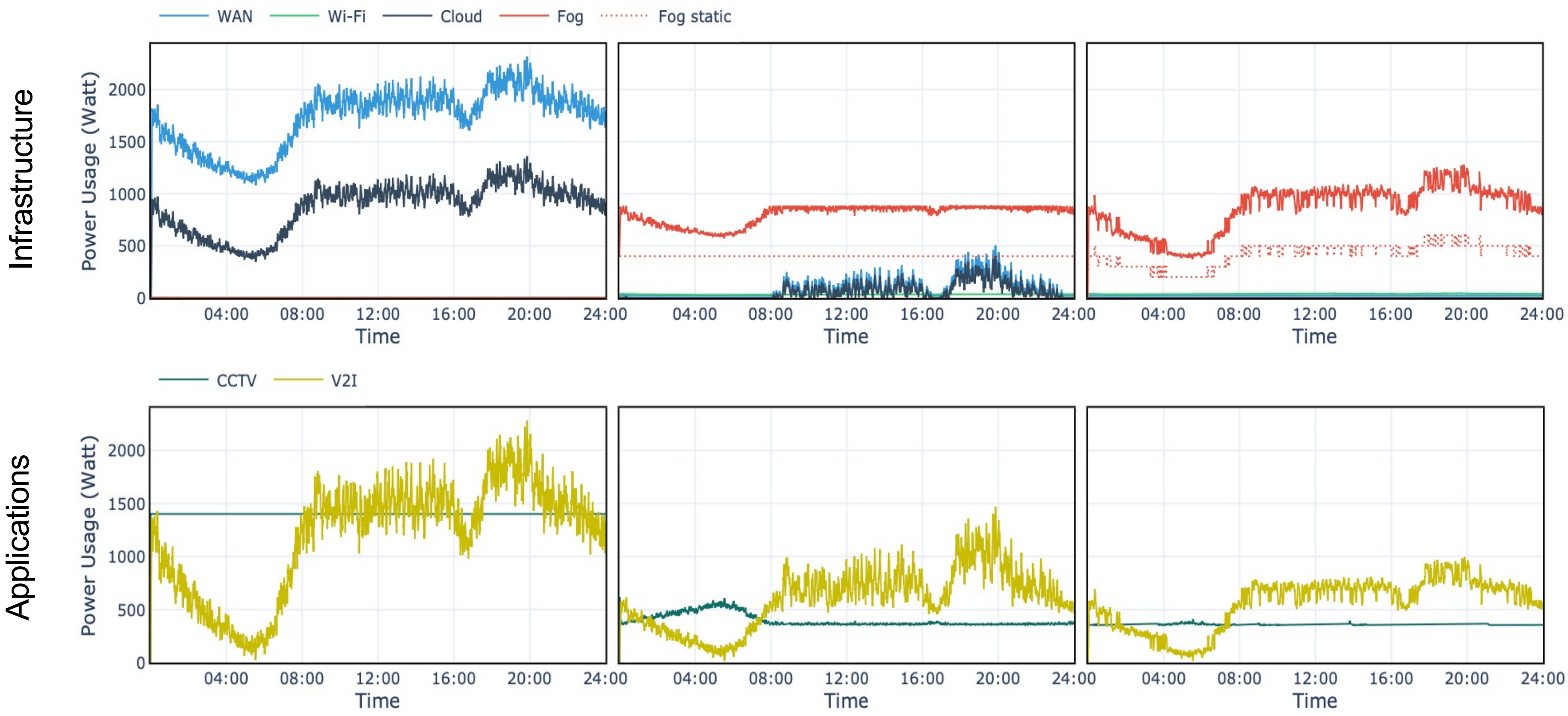


Two types of applications:
- CCTV: One for each traffic light
- V2I: One for each taxi

Experiment Results



Results Over Time



cloud only

4 fog nodes

6 fog nodes with energy saving mechanism

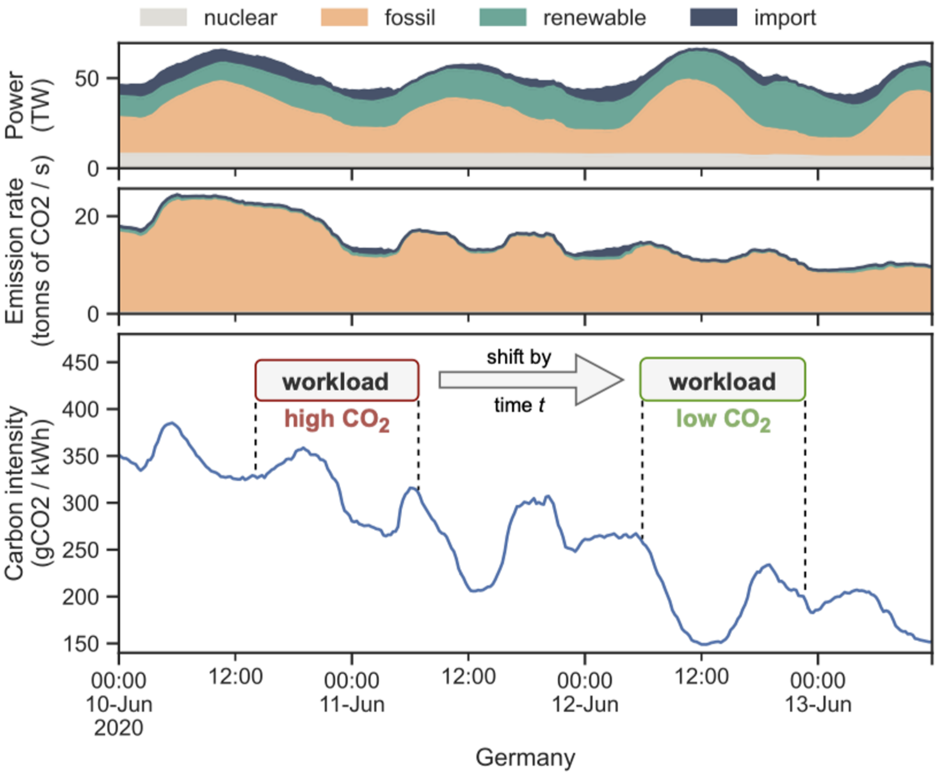
Cloud Workload Shifting

Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud. Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. 22nd ACM/IFIP International Middleware Conference (Middleware). ACM. 2021.

Motivation

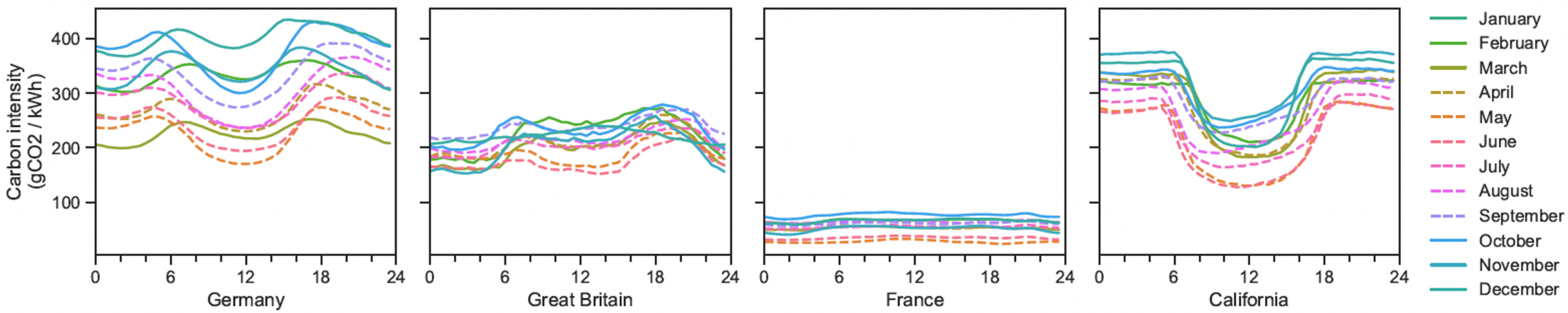
- Data centers are responsible for >1% of global energy consumption, with best case projections for 3% in 2030 [1, 2]

- Low-carbon objective: Compute when and where low-carbon energy is available



Changing Carbon Intensity

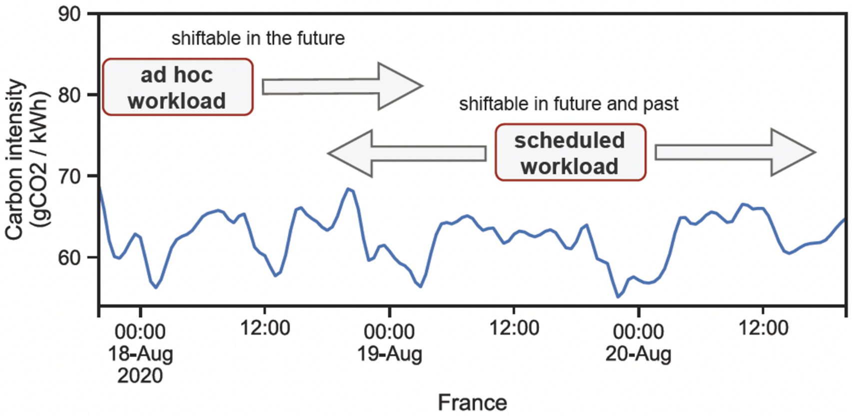
- What are the most promising times to shift work to?

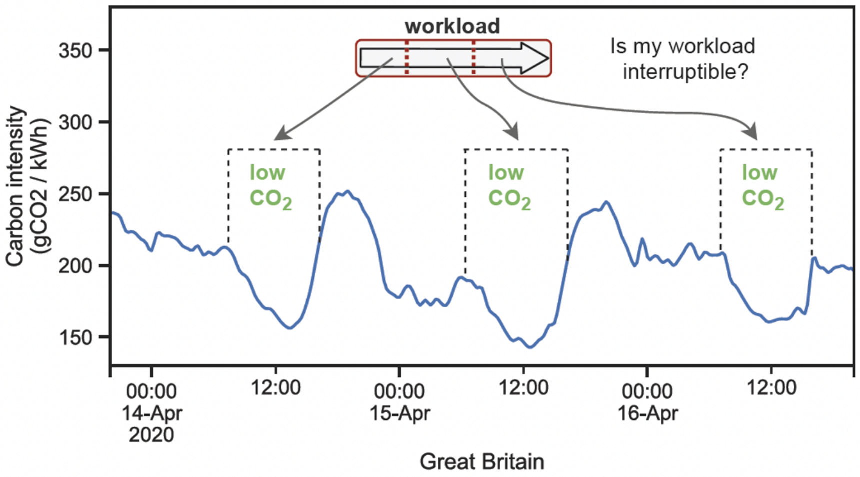


- Carbon intensity: Amount of CO2 equivalent greenhouse gases emitted per kilowatt hour of energy

Shiftable Cloud Workloads

- Ad hoc vs. scheduled
- Temporal constraints
- Job runtimes
- Interruptibility





Experimental Evaluation

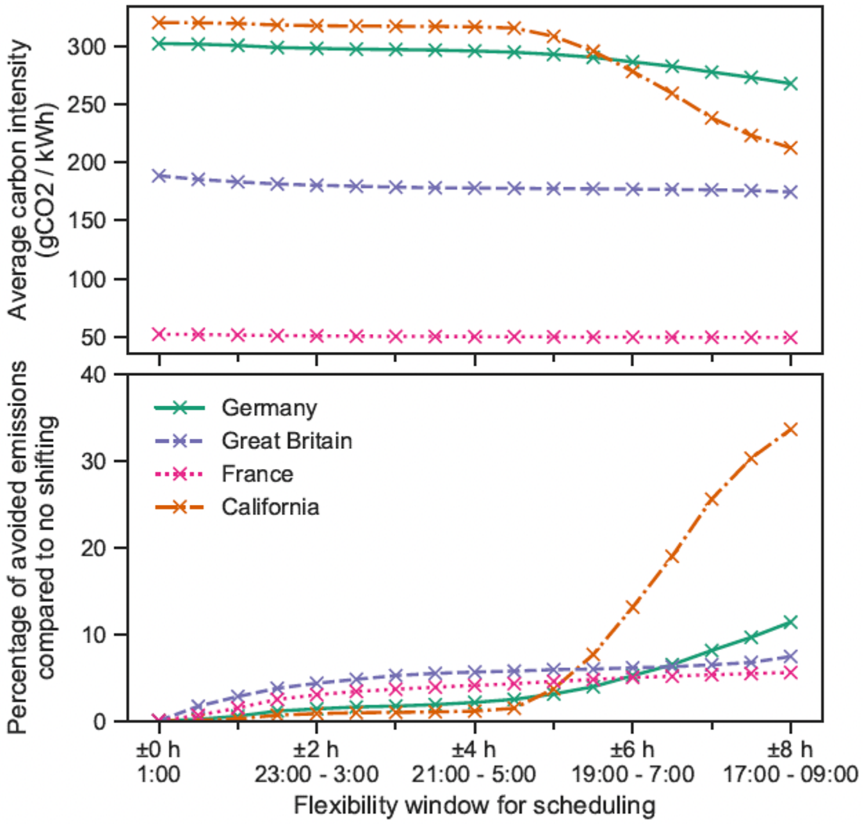
- Evaluation of two scenarios using the LEAF simulator:

- Scenario 1 – Periodic Jobs: Nightly builds, integration tests, database backups, generation of business reports, …

- Scenario 2 – Ad Hoc Jobs: ML training jobs, CI/CD, data analysis pipelines, scientific simulations, …

Scenario 1: Periodic Jobs

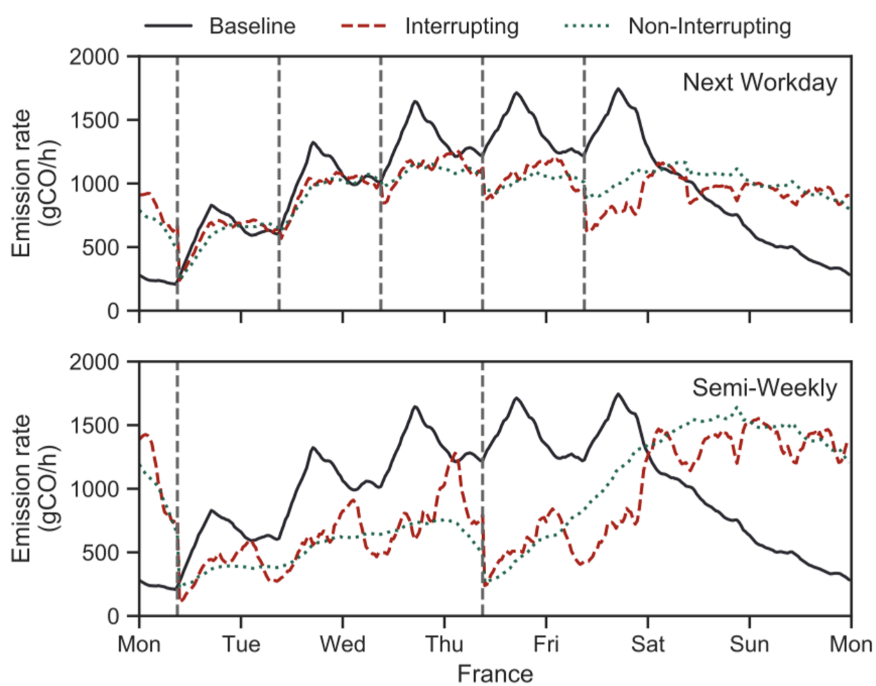
- Baseline: All jobs scheduled at
  1 am in the night
- Increasing the window by +- 1h to allow scheduling between
  - 00:00 to 3:00 (+- 1h)
  - 23:00 to 4:00 (+- 2h)
  - …
  - 17:00 to 9:00 (+- 8h)



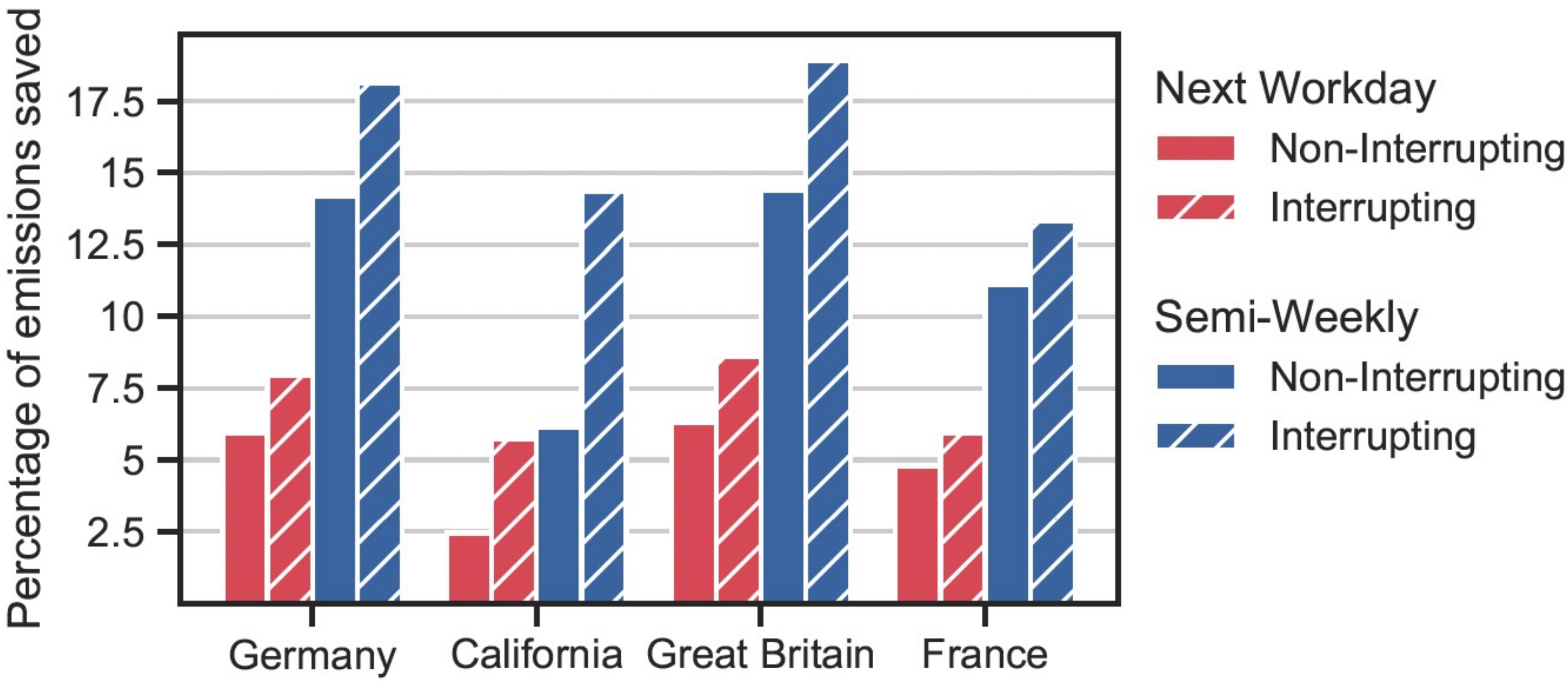
Scenario 2: Large Ad Hoc Jobs

- Setting: Jobs arrive randomly during working hours (Mo - Fr, 9:00h - 17:00h)

- Baseline: Instant scheduling
- Investigate influence of
  - Deadlines
  - Interruptibility



Scenario 2: Overall Results



Admission Control

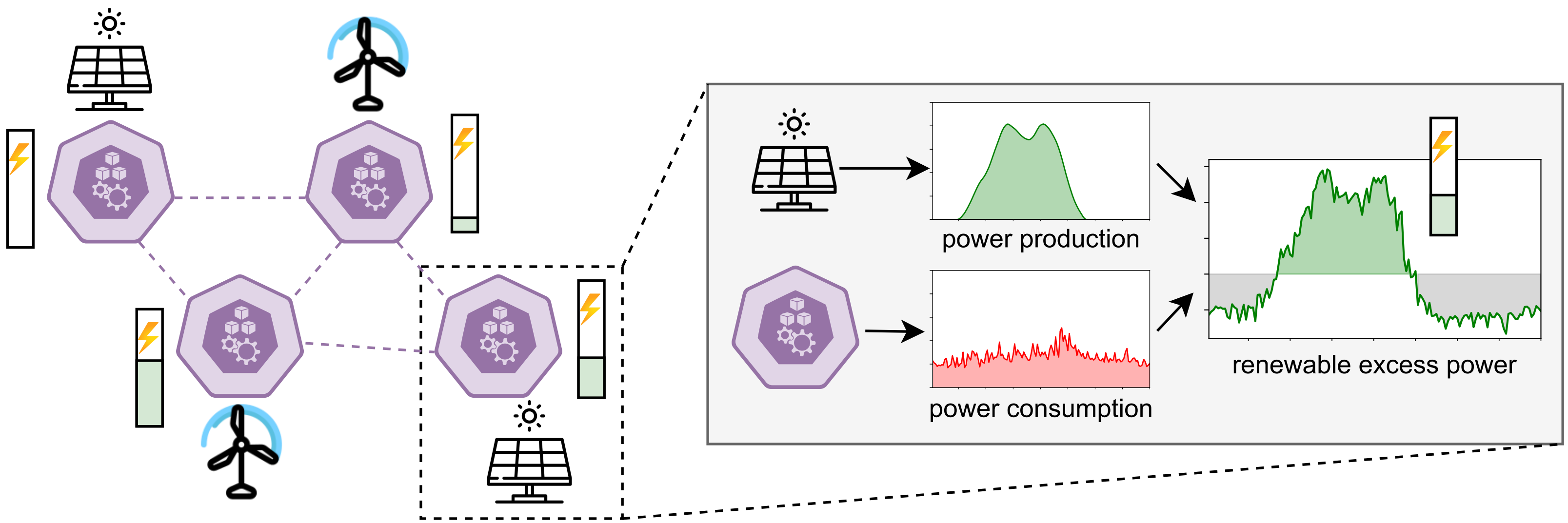
Cucumber: Renewable-Aware Admission Control for Delay-Tolerant Cloud and Edge Workloads. Philipp Wiesner, Dominik Scheinert, Thorsten Wittkopp, Lauritz Thamsen, and Odej Kao. Currently under review. 2022.

Problem Setting

- Assumptions:
  - Access to renewable energy sources, with not all the energy being used always, yet also no energy storage
  - Resource constrained compute nodes running a high-priority, time-critical base load, but over time also spare resources

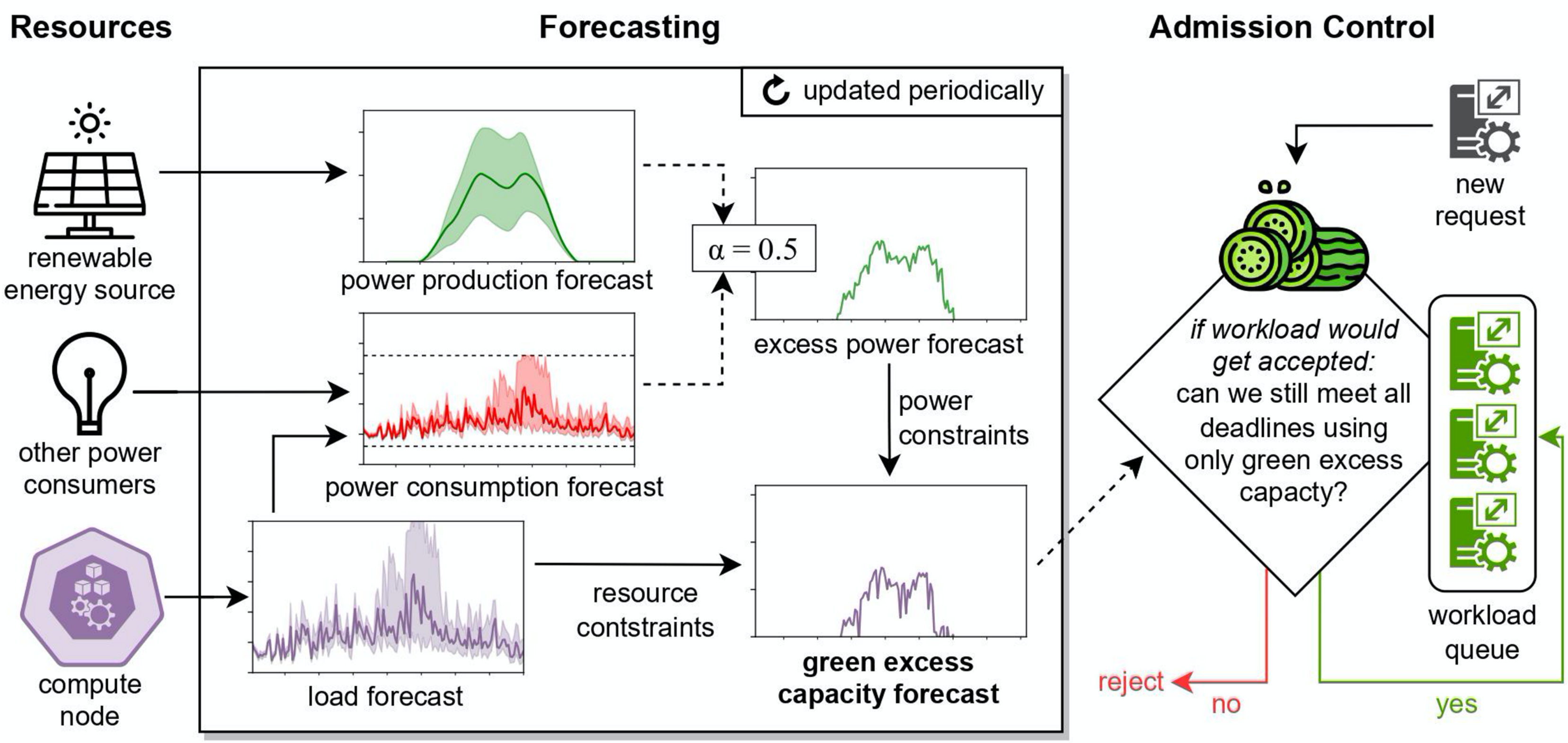
- Goal: Utilize excess energy by computing low-priority, delay-tolerant workloads

Renewable Excess Power



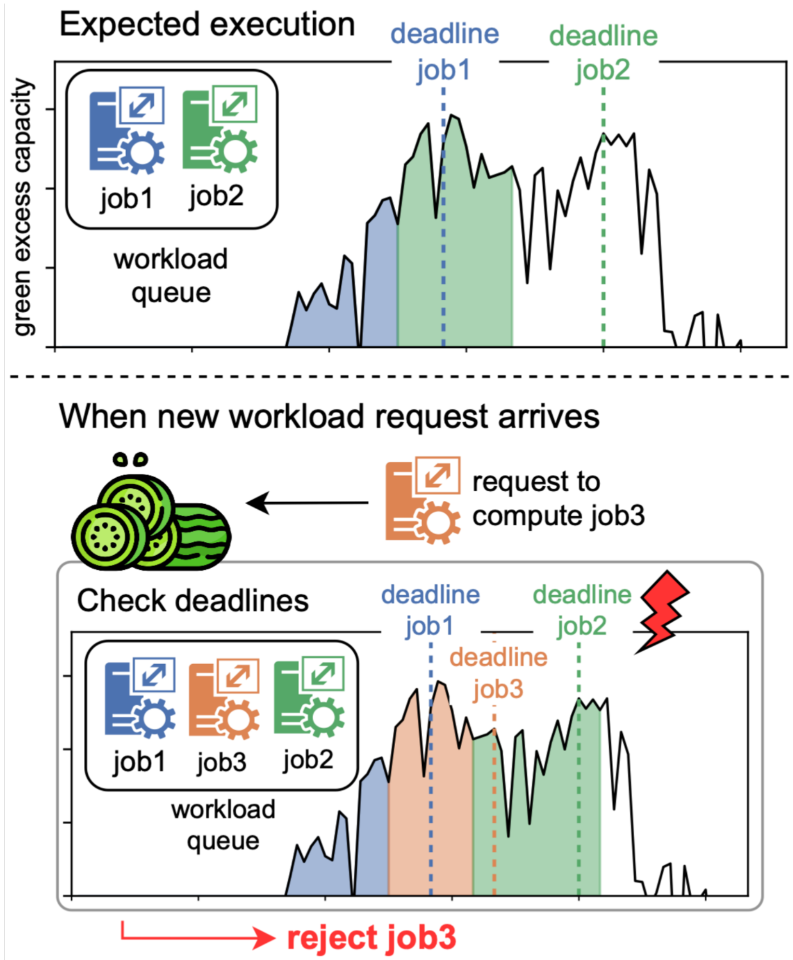
- In some settings local demand does temporarily not cover all produced power, yet is also not put into storage or grids

The "Cucumber" Concept

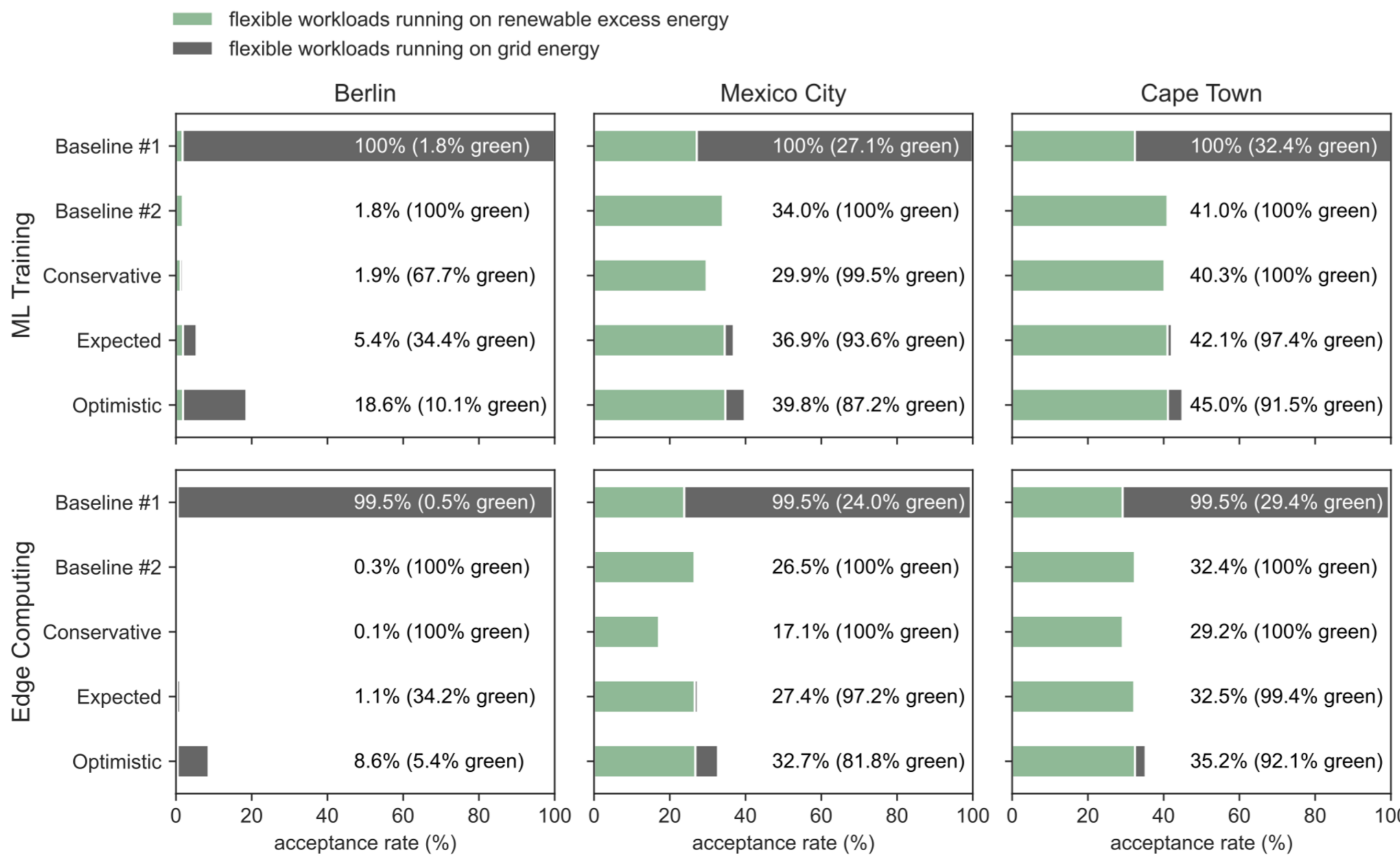


Admission Control

- For each incoming job, Cucumber checks if it can be computed using excess energy only
- Through probabilistic forecasts, admission can be tuned towards
  - conservative (low acceptance rate, low grid power usage) or
  - optimistic results (vice versa)

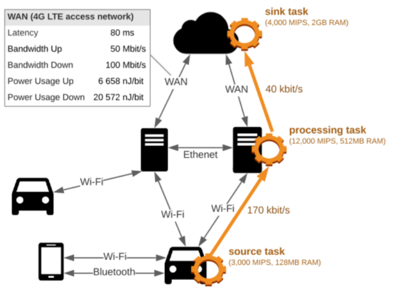


Evaluation Results



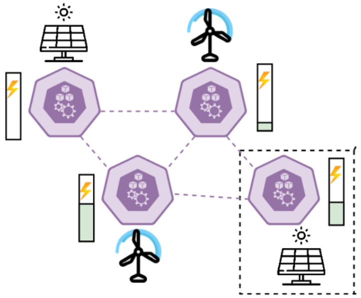
Outlook & Summary

Ideas for Next Steps



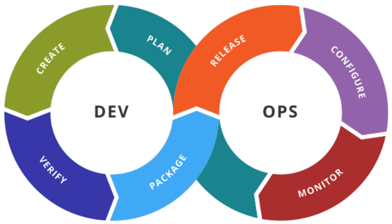
”Truer-to-life” experiments:

- More realistic and interesting simulations?
- Experiments with real hardware and systems?



Decentralized resource management:

- What happens beyond one node / datacenter?
- How should nodes negotiate workloads?



Continuous feedback for developers:

- How can CI/CD tools also report the energy consumption and emissions of applications?

Summary

- Taking carbon emission into account when managing edge/cloud computing workloads could reduce emissions
- Simulation, forecasting, and optimization methods will be valuable tools
- Not as clear how to get this into real services, systems, and developer tools though
- Interesting research into distributed systems, resource management, and also software engineering ahead

References

[1] Eric Masanet, Arman Shehabi, Nuoa Lei, Sara Smith, and Jonathan Koomey. “Recalibrating Global Data Center Energy-Use Estimates”. Science 367. 2020

[2] Anders S.G. Andrae and Tomas Edler. “On Global Electricity Usage of Communication Technology: Trends to 2030”. MDPI Challenges 6(1). 2015

Philipp Wiesner and Lauritz Thamsen. “LEAF: Simulating Large Energy-Aware Fog Computing Environments”. In *5th IEEE International Conference on Fog and Edge Computing (ICFEC).* 2021

Philipp Wiesner, Ilja Behnke, Dominik Scheinert, Kordian Gontarska, and Lauritz Thamsen. “Let's Wait Awhile: How Temporal Workload Shifting Can Reduce Carbon Emissions in the Cloud”. In 22nd ACM/IFIP International Middleware Conference (Middleware). ACM. 2021