Reducing the impact of energy consumption from computing with CATS, The **C**limate **A**ware **T**ask **S**cheduler

Sadie Bartholomew
University of Reading (Dept. Meteorology), National Centre for Atmospheric Science (NCAS)

On behalf of the full CATS team:
Colin Sauzé, Abhishek Dasgupta, Andrew Walker, Loïc Lannelongue, Thibault Lestang, Tony Greenberg, Lincoln Colling, Adam Ward and Carlos Martinez

For Computing Insight UK 2024 (Sustainability Strand), 2024-12-05

SOFTWARE SUSTAINABILITY INSTITUTE

National Centre for Atmospheric Science
NATURAL ENVIRONMENT RESEARCH COUNCIL

University of Reading

Border image credits: 'Climate Stripes' infographic designed by Prof. Ed Hawkins (University of Reading), see showyourstripes.info

Motivating question

Image credits: https://i.imgflip.com/208mpa.jpg, from IT Crowd (Channel 4)

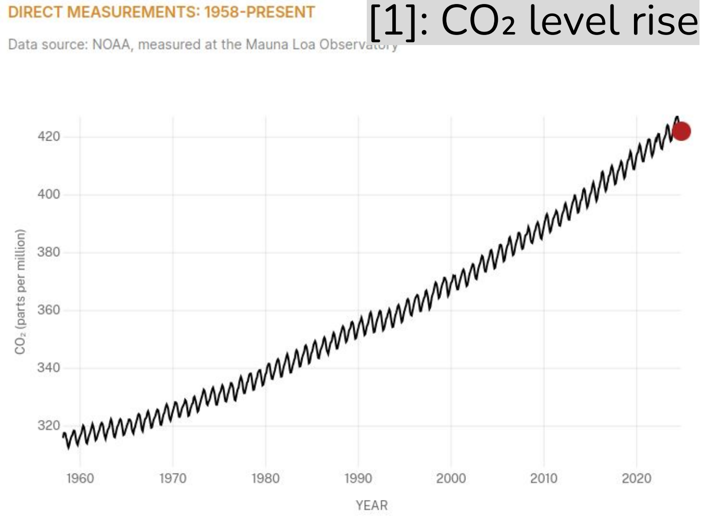
Computing *always requires energy* (electricity etc.) - how can we do it in a sustainable way to not exacerbate the climate crisis?

Motivating issue: the climate crisis

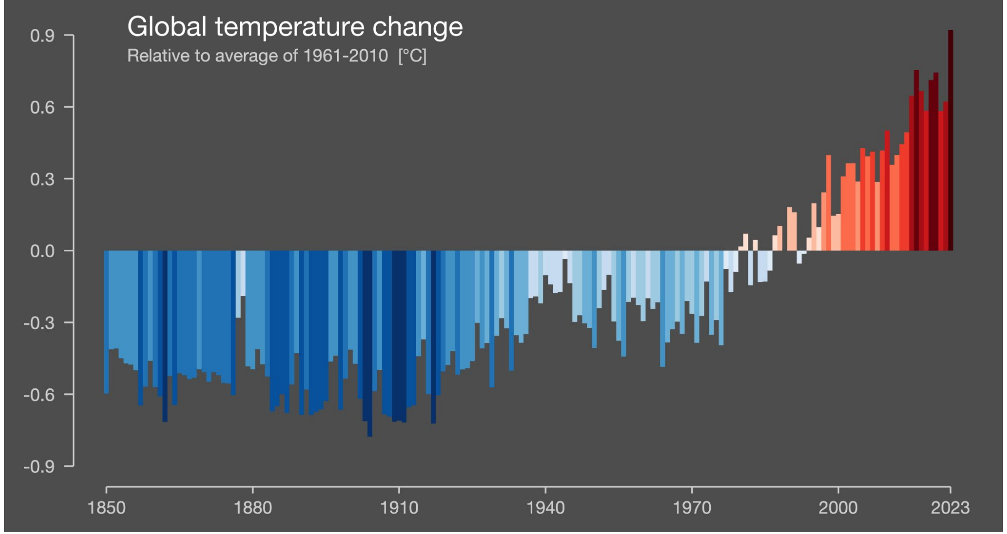
Human activities, notably fossil fuel burning to generate energy, are (largely) responsible for accumulation of greenhouse gases (e.g. CO_2 [1]) in the Earth’s atmosphere causing rise in global temperatures [2] and sea levels [3] etc. - **activities including computing**

Sources of plots:
showyourstripes.info,
https://climate.nasa.gov/vital-signs/

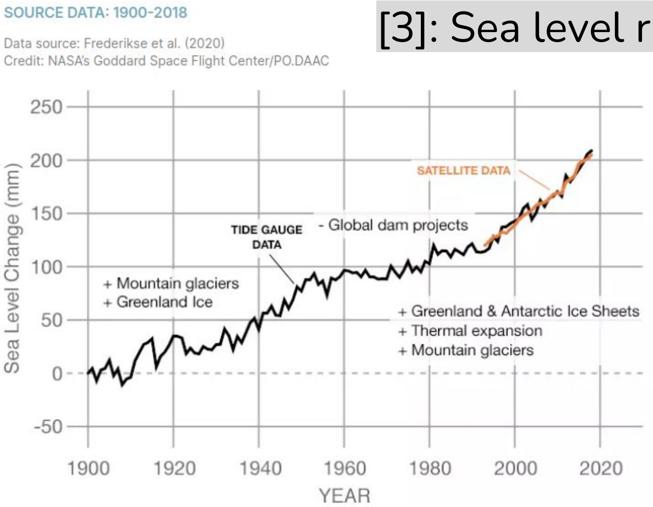
[1]: CO_2 level rise



[2]: Atmospheric temperature rise



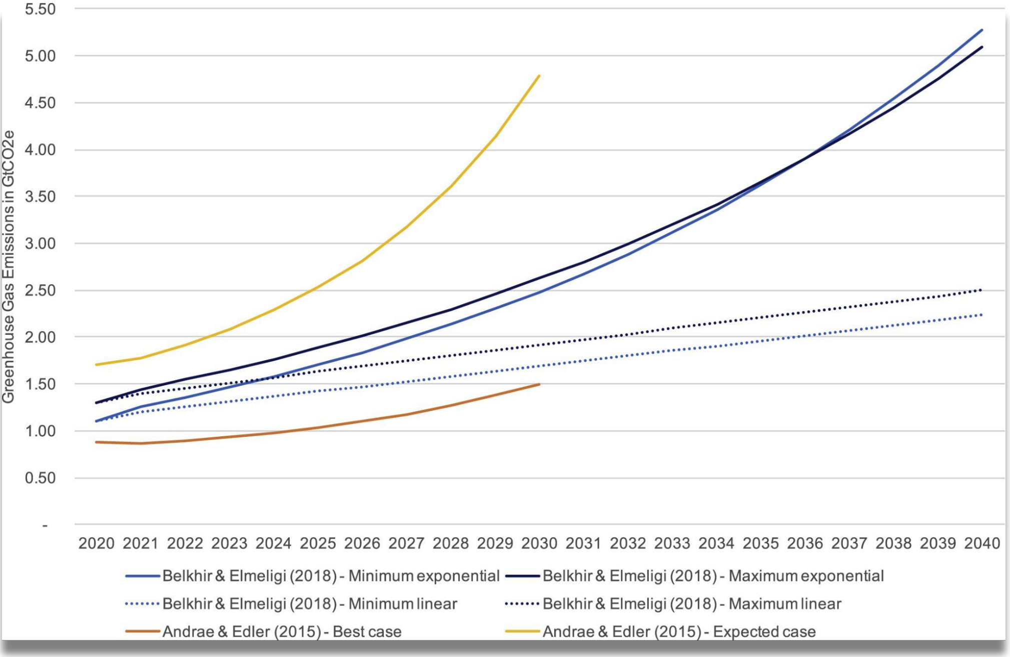
[3]: Sea level rise



The consumption of computing: significant & increasing!

- The ICT sector uses 4-10% of the world's electricity and generates 1.5-5% of its greenhouse gas emissions*

- Computing's global share is modest but growing at a much faster rate than many other energy-consuming sectors. Energy demands of data centers and HPC systems are expected to increase significantly over the next few decades, driven partly by growing use of cloud services, AI, and machine learning models plus more need for data centers which are power hungry



- “ICT's footprint has likely grown faster than global emissions, with a very uncertain best estimate of twice as fast”†, though it is hard to estimate accurately (see plot†)

*Source: The EU climate strategy for the ICT sector

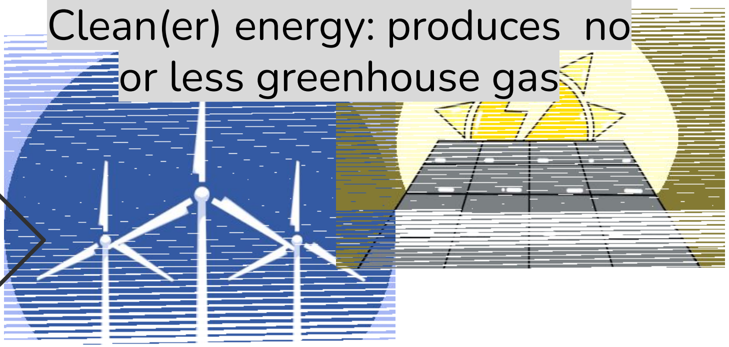
† Source: ‘The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations’, Freitag et al., https://doi.org/10.1016/j.patter.2021.100340

The underlying idea

We can (and should) work to reduce our energy consumption from the computing we do. But we can *also* reduce our climate impact by being *more clever* with the *set energy we do use* so that we end up using **more energy from renewables (clean) than fossil fuels (dirty)**

Dirty energy: produces (more) greenhouse gas

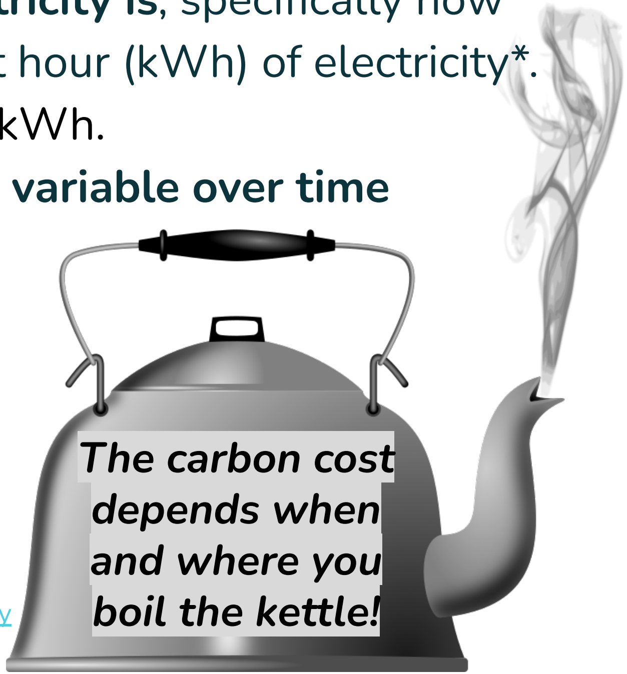
Aim to shift our energy consumption this way



Clean(er) energy: produces no or less greenhouse gas

How do we measure the ‘cleanliness’ of energy we use?

- Because renewable sources aren’t available in a steady manner, and demand on a given electricity grid varies, CO_2 emissions from some task requiring a set amount of electricity *depend on the datetime and location in which the task is done*
- **Carbon intensity is a measure of how clean our electricity is**, specifically how many grams of CO_2 are released to produce a kilowatt hour (kWh) of electricity*. This becomes our metric of interest. Units are gCO_2e/kWh.
- The **carbon intensity of electricity (in the UK) is very variable over time**
  - Windy and/or sunny weather ⇒ lower carbon
  - Generally between 0 and 400 gCO_2e/kWh
  - EU average 251 gCO_2e/kWh in 2022[†]



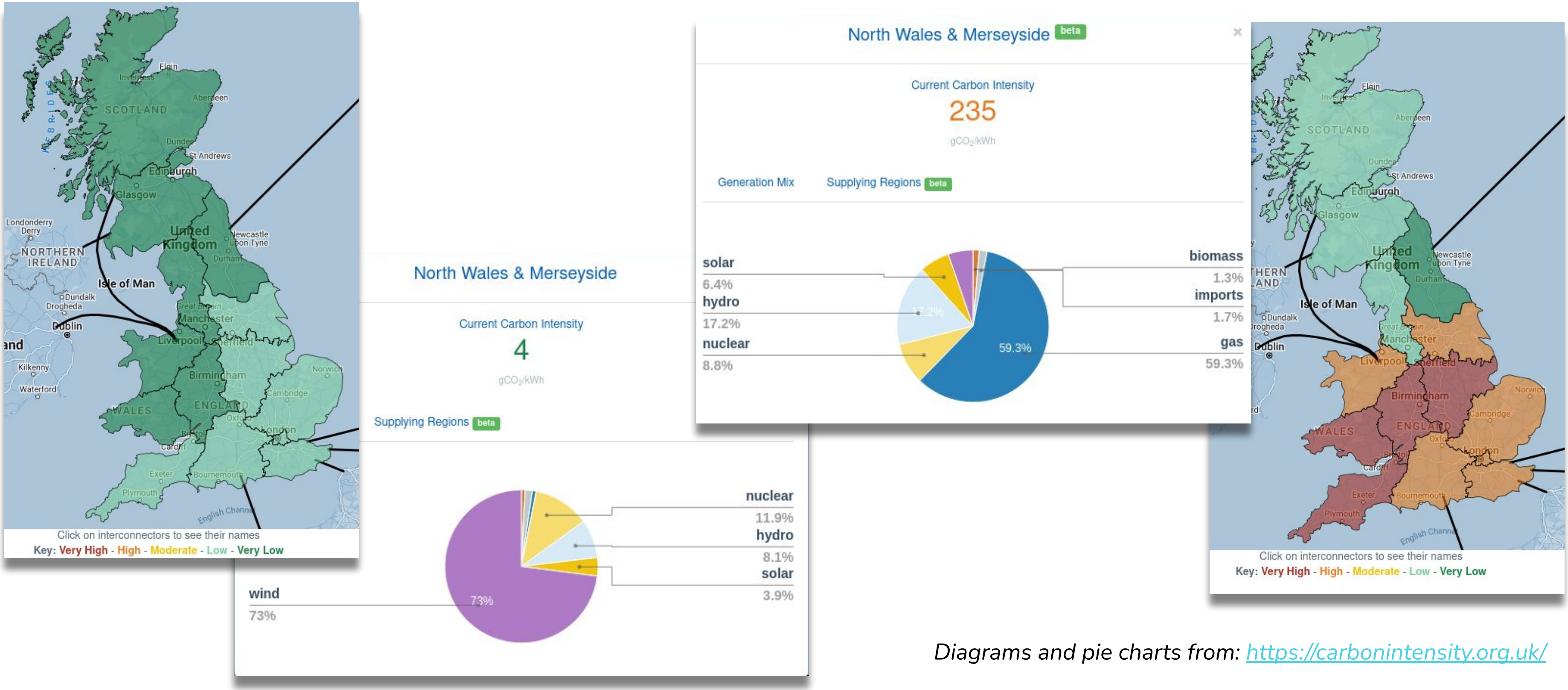
The carbon cost depends when and where you boil the kettle!

* *Source of definition*: https://www.nationalgrid.com/stories/energy-explained/what-is-carbon-intensity

[†] https://www.eea.europa.eu/en/analysis/maps-and-charts/co2-emission-intensity-15

Regional & weather-based influence on carbon intensity

- Left: windy & quite sunny day across UK, right: neither windy nor sunny across UK
- For example showing regional carbon intensity factors for North Wales & Merseyside



North Wales & Merseyside

Current Carbon Intensity

4

gCO$_2$/kWh

Supplying Regions beta

North Wales & Merseyside beta

Current Carbon Intensity

235

gCO$_2$/kWh

Generation Mix Supplying Regions beta

Click on interconnectors to see their names

Key: Very High - High - Moderate - Low - Very Low

Click on interconnectors to see their names

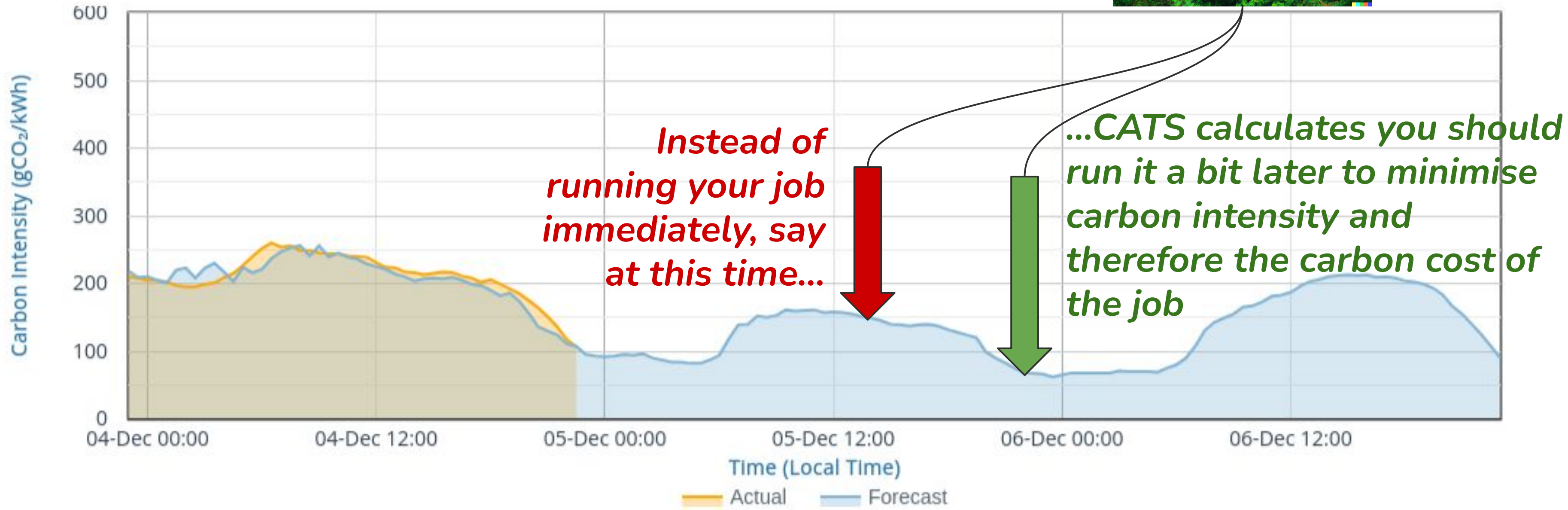
Key: Very High - High - Moderate - Low - Very Low

Diagrams and pie charts from: https://carbonintensity.org.uk/

Introducing our tool CATS to manage time-shifting of jobs

The **C**limate **A**ware **T**ask **S**cheduler (https://github.com/GreenScheduler/cats) calculates the optimal time to run a job to minimise its carbon intensity

Carbon Intensity Forecast (-24hrs to +48hrs)



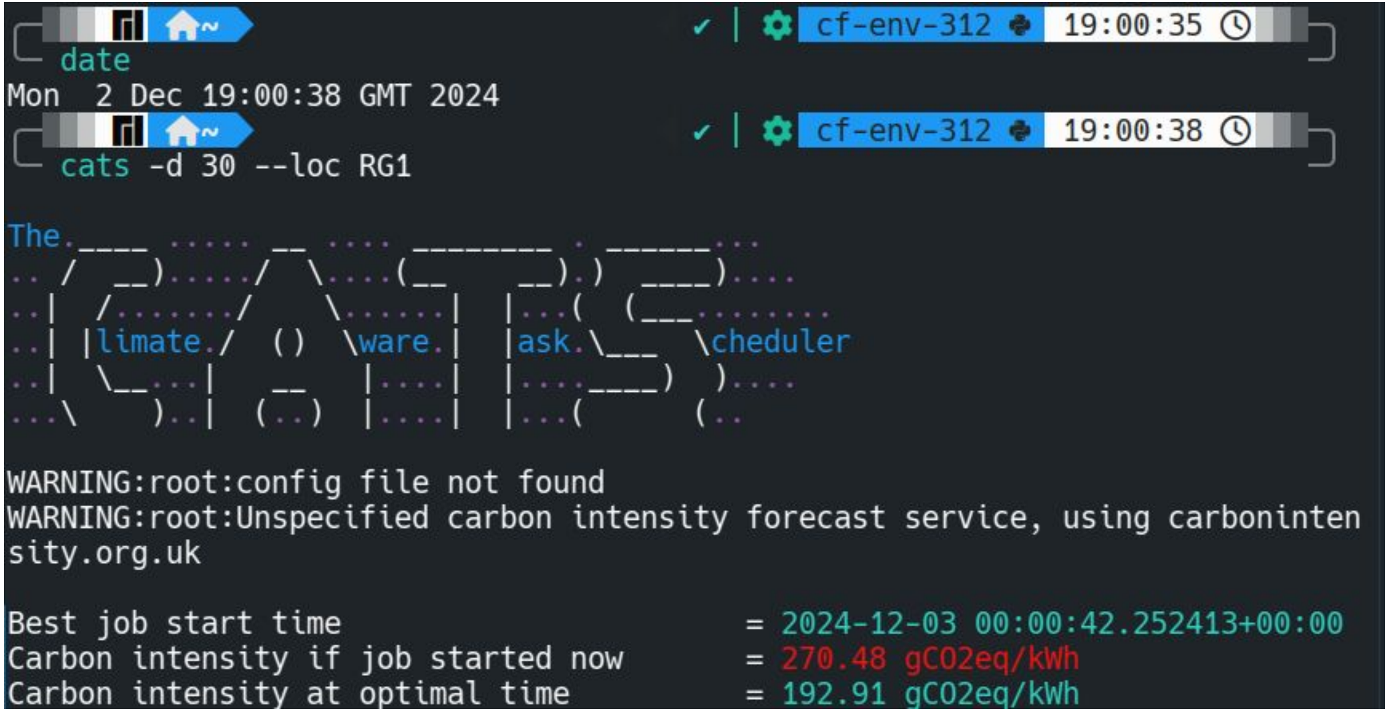
Instead of running your job immediately, say at this time...

...CATS calculates you should run it a bit later to minimise carbon intensity and therefore the carbon cost of the job

Plot from: https://carbonintensity.org.uk/ (more on this resource in later slides!)

Basic usage of CATS: via a command-line interface

- Terminal use, with configuration via YAML file and/or CLI options & arguments
- Minimal use: `cats -d <job duration in mins> --loc <postcode>`
- Example: shows savings of >75 gCO2e/kWh by waiting ~6h to run a job



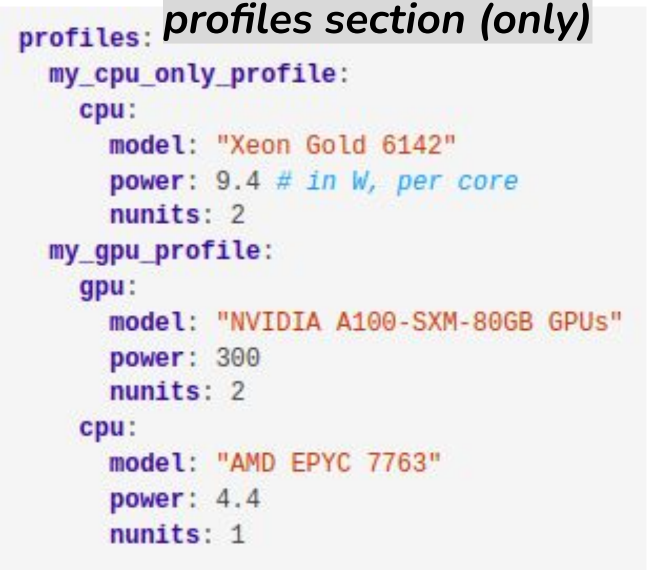
```
  ~                                               ✓ |  cf-env-312   19:00:35
  date
Mon  2 Dec 19:00:38 GMT 2024
  ~                                               ✓ |  cf-env-312   19:00:38
  cats -d 30 --loc RG1

The.____ ..... __ .... ________ . ______...
.. /  __)...../  \....(__    __).)  ____)....
..|  /......./    \......|  |...(  (___........
..| |limate./  ()  \ware.|  |ask.\___  \cheduler
..|  \__...|   __   |....|  |....____)  )....
...\    )..|  (..)  |....|  |...(      (..

WARNING:root:config file not found
WARNING:root:Unspecified carbon intensity forecast service, using carboninten
sity.org.uk

Best job start time                       = 2024-12-03 00:00:42.252413+00:00
Carbon intensity if job started now       = 270.48 gCO2eq/kWh
Carbon intensity at optimal time          = 192.91 gCO2eq/kWh
```

Further usage: direct scheduling & estimating carbon footprint

- **To directly schedule a job with the CATS calculation**, use the argument `--scheduler`. We currently support the UNIX `at` command, for example to run a Python script `work.py` expected to take an hour or so: `cats -d 60 --loc RG1 --scheduler at --command 'python work.py'`

- **You can go further than carbon *intensity* information and extract the estimated carbon *footprint* reduction** from delaying the compute if you provide memory consumption and a hardware profile for the relevant machine: `cats --duration 480 --location "EH8" --footprint --memory 16 --profile my_gpu_profile --gpu 4 --cpu 1`

Example YAML config file, profiles section (only)



```yaml
profiles:
  my_cpu_only_profile:
    cpu:
      model: "Xeon Gold 6142"
      power: 9.4 # in W, per core
      nunits: 2
  my_gpu_profile:
    gpu:
      model: "NVIDIA A100-SXM-80GB GPUs"
      power: 300
      nunits: 2
    cpu:
      model: "AMD EPYC 7763"
      power: 4.4
      nunits: 1
```

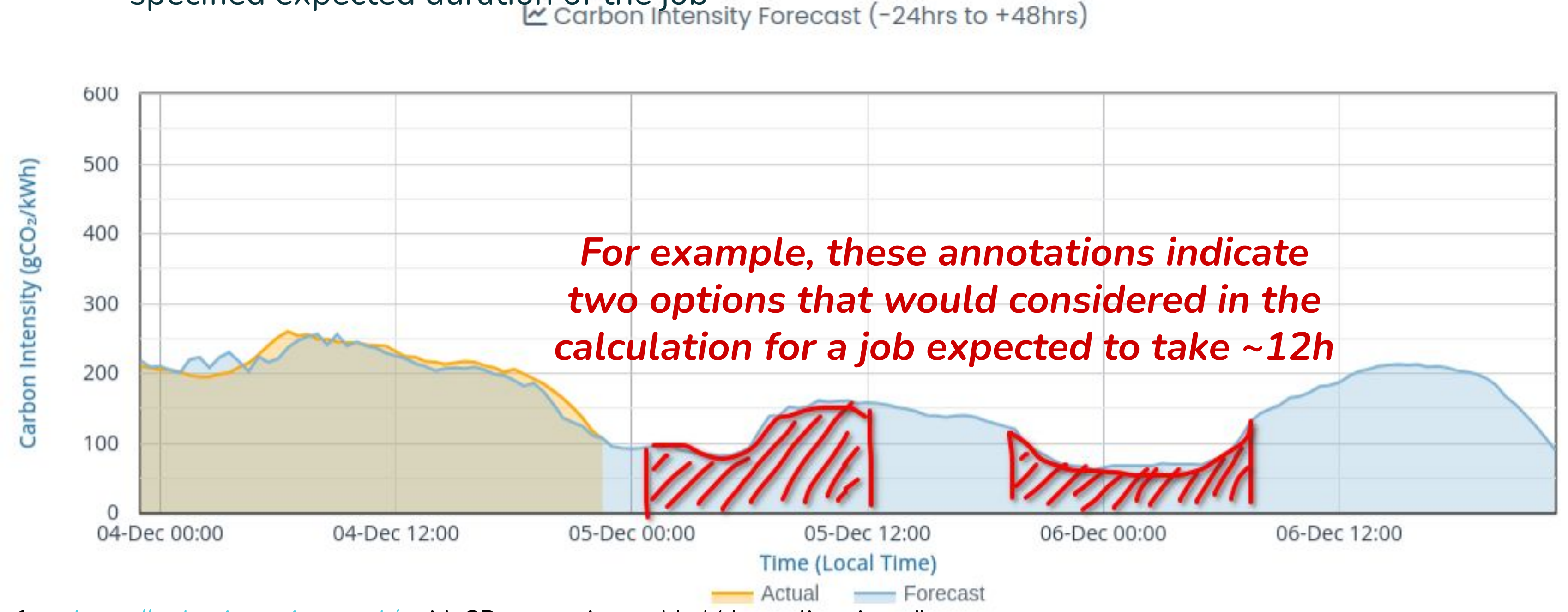
A brief history of CATS

May 2023 (SSI CW23) → ***The future***

- Devised & prototyped at the Software Sustainability Institute’s Collaborations Workshop 2023 Hack Day (winning first prize!), proof of concept intended for small-scale compute
- Original hackathon team took the project forward together to continue developing CATS
- Version 1.0 released in July this year, marking the first release of a stable tool (full documentation, improved CLI, test coverage & output formats for humans & machines)
- Work in progress with further support from the SSI, including integration with SLURM, testing on real HPCs & submitting a publication to the Journal of Open Source Software

How does CATS work?

- To run software when renewable sources of energy are most plentiful, CATS:
  - uses National Grid ESO's Carbon Intensity API (carbonintensity.org.uk) for carbon intensity forecast
  - takes such data appropriate to the local region (found from a given postcode as proxy for location)
  - calculates to effectively minimise the area under the curve (as illustrated on the plot here) for the specified expected duration of the job



For example, these annotations indicate two options that would considered in the calculation for a job expected to take ~12h

Plot from *https://carbonintensity.org.uk/*, with SB annotations added (drawn lines in red)

Use cases for CATS: from small- to large-scale compute

- **Version 1.0** (July 2024): first stable/mature release, designed for ‘small-scale’ computing e.g. a few hours on a workstation/desktop or laptop overnight

- Work towards **Version 2.0** is in progress, which aims to target the more pressing source of carbon emissions, HPC and HTC
    - Includes work to test CATS on a ‘mini HPC’ (Raspberry Pi cluster, funded by the SSI and built by CATS team members Sadie and Colin)

Further work and upcoming version 2!

- Work underway for integration with the batch scheduler SLURM (https://slurm.schedmd.com/) which will be in CATS version 2
  - Simplest approach: using `sbatch` to offset start time
  - Our ideal result: HPC systems can implement ‘green’ queues to use CATS to delay jobs that users are happy to in return for reduced carbon footprint (and/or incentives)
  - Integrating carbon accounting as a Slurm plugin (will need rewrite in C)
  - SSI funding provided a few months of developer time, coming to the end of this and approaching completion of work

Example of v2 carbon footprint saving for a fictional HPC

- For an example of a fictional HPC, with hardware as follows:
  - 64 core AMD EPYC 7773X (Milan) CPUs
    - 10 nodes, 2 CPUs per node, 20 CPUs total, 1280 cores
  - Fully loaded CPU = 255 W, Idle CPU = 37.5 W (from https://www.phoronix.com/review/amd-epyc-7773x-linux/9)
  - Idle saving = 217.5 W per CPU
  - Cluster idle vs peak = 4.35 kW
- Time shifting reduces grid intensity from 200 to 50 g/kWh = 150 g/kWh reduction
- The calculation:
  - 12 hour job using all cores
  - 12h * 4.35 kW = 52.2 kWh
  - 52.2 kWh * 0.15 kg = 7.83 kg
- Comparable to driving an average car (150 g/km) 50 km (7.5 kg)!

Limitations of CATS and notes regarding value

- Only works for the UK (at present) due to lack of APIs like the National Grid ESO's Carbon Intensity one used, for other countries/regions (open Issue https://github.com/GreenScheduler/cats/issues/22)
- Relies on user specifying the job length correctly - and this can be hard to estimate and might require pre-run(s) to estimate well (enough)
- Won’t be able to do much on systems at/near 100% load
- Can’t handle jobs expected to take more than 2 days due to forecast cutoff of the National Grid ESO API
- *Note*: the UK electricity grid is planned to be net-zero by 2035, but that’s quite optimistic and besides, if we can do something now, then why wait?
- *Note*: not the only thing you can/should do to reduce the climate impact of your computing! You can look to also reduce emissions from scope 3 (manufacturing), cooling, storage & networks (e.g. see blog post ‘Tracking the environmental impact of research computing’ SSI blog post covering useful background: https://www.software.ac.uk/blog/tracking-environmental-impact-research-computing)

Summary of CATS, The Climate Aware Task Scheduler

- Computing uses (a lot of!) energy - HPC, HTC, data centers and AI in particular
- One approach to reduce our impact on the climate crisis from greenhouse gas emissions resulting from our energy consumption is to shift to **using more of the ‘clean’ renewable sources over ‘dirty’ sources like fossil fuels**
- We can do this by **using the local electricity when it is lower in carbon intensity**
- By **intelligently time shifting compute jobs** to run them at the time that minimises carbon footprint across their expected duration, using real-time carbon intensity data from the National Grid ESO API, CATS can contribute to more sustainable computing
- CATS was **initially developed for small-scale compute jobs, but work is underway to support HPC/HTC** (better targets for reducing carbon impact!) via SLURM integration
- For now, **try out CATS Version 1**! See https://github.com/GreenScheduler/cats

Border image credits: ‘Climate Stripes’ infographic designed by Prof. Ed Hawkins (University of Reading), see showyourstripes.info

Thanks for listening.

For more info. about CATS and/or other aspects from this talk, please ask me anything now or you can explore such resources as:

- the CATS codebase, OSS on Github: https://github.com/GreenScheduler/cats
- the CATS package documentation: https://greenscheduler.github.io/cats/
- a recent episode of the ‘Code for Thought’ podcast in which myself and Colin talk about CATS: https://www.buzzsprout.com/1326658/episodes/15766448-en-bonus-green-computing-at-the-rse-conference-2024-in-newcastle?t=0
- ‘Tracking the environmental impact of research computing’ SSI blog post covering useful background: https://www.software.ac.uk/blog/tracking-environmental-impact-research-computing

Border image credits: ‘Climate Stripes’ infographic designed by Prof. Ed Hawkins (University of Reading), see showyourstripes.info