



Green Cloud Computing

University of Manchester Bohr Seminar,
Joshua Horswill

Who am I and what do I do?



- Data scientist at Greenpixie, LHCb/University of Manchester alumn.
- Worked on optimising the first stage of the LHCb trigger.
- Analysed CP asymmetry in beauty to open charm decays.
 - Now I research the impact of cloud compute (with a focus on AI) and build data pipelines to serve customers with the results of this research.



What does Greenpixie do?



- Greenpixie appends sustainability data to cloud and software-as-a-service billing reports. This enables:
 - Accurate reporting and disclosure for Greenhouse Gas protocol compliance.
 - Improving frugality when provisioning resources in the cloud
 - The cloud-native sustainability tools leave a lot of people wanting, so we fill that gap

Product Code	Line Item Usage Type	Instance Type	Pricing Unit	Region	Energy (kWh)	Carbon (tCO ₂ e)	Water (L)	Estimate Category
AmazonEC2	BoxUsage	m1.small	Hrs	us-east-1	0.0068683	3.526e-6	0.017096	compute
AmazonEC2	BoxUsage:a1.2xlarge	a1.2xlarge	Hrs	us-east-1	0.014708	9.685e-6	0.036611	compute
AmazonEC2	BoxUsage:a1.4xlarge	a1.4xlarge	Hrs	us-east-1	0.029416	1.937e-5	0.073221	compute
AmazonEC2	BoxUsage:a1.large	a1.large	Hrs	us-east-1	0.003677	2.421e-6	0.0091526	compute
AmazonEC2	BoxUsage:a1.medium	a1.medium	Hrs	us-east-1	0.0018385	1.211e-6	0.0045763	compute
AmazonEC2	BoxUsage:a1.metal	a1.metal	Hrs	us-east-1	0.029416	1.937e-5	0.073221	compute
AmazonEC2	BoxUsage:a1.xlarge	a1.xlarge	Hrs	us-east-1	0.007354	4.843e-6	0.018305	compute
AmazonEC2	BoxUsage:c1.medium	c1.medium	Hrs	us-east-1	0.014024	7.136e-6	0.034908	compute
AmazonEC2	BoxUsage:c1.xlarge	c1.xlarge	Hrs	us-east-1	0.058779	3.012e-5	0.14631	compute

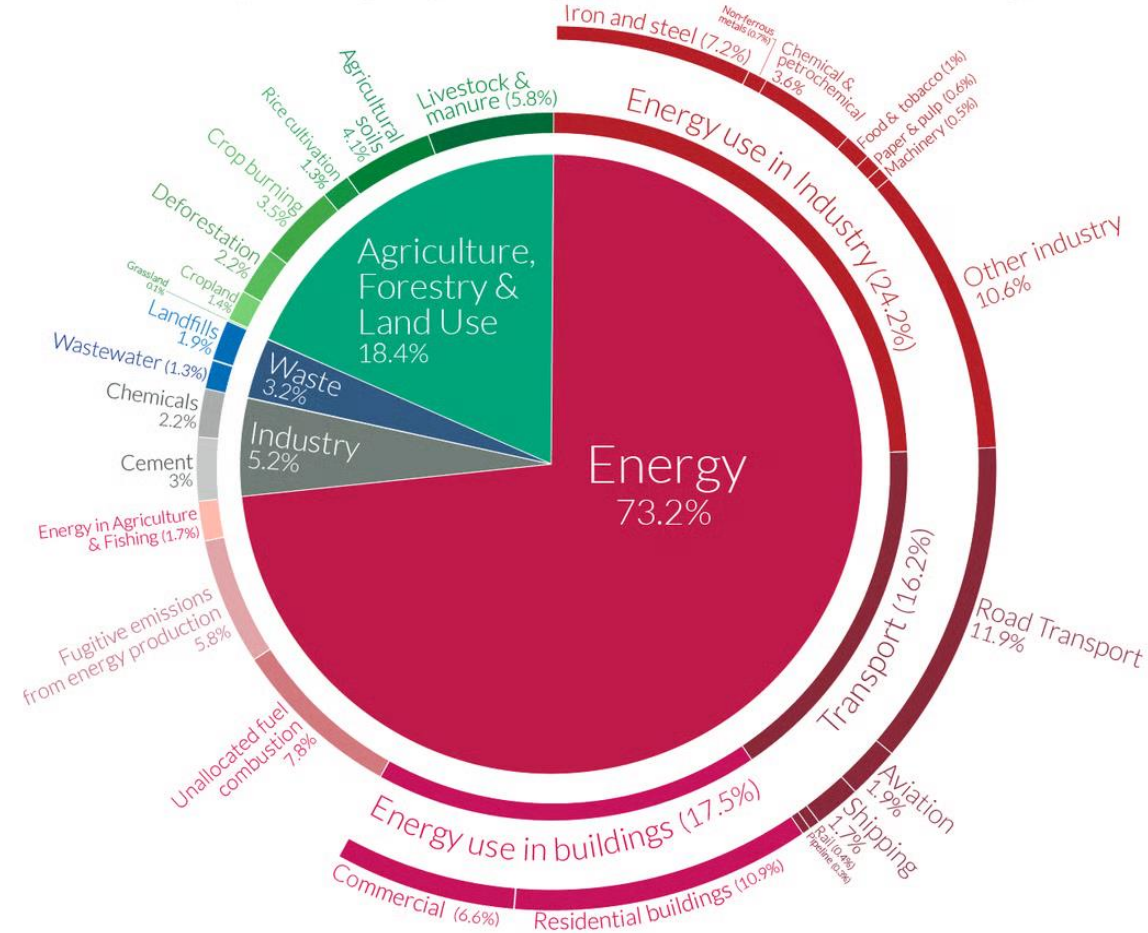
CO₂-eq emissions from the ICT sector are larger than the aviation industry.



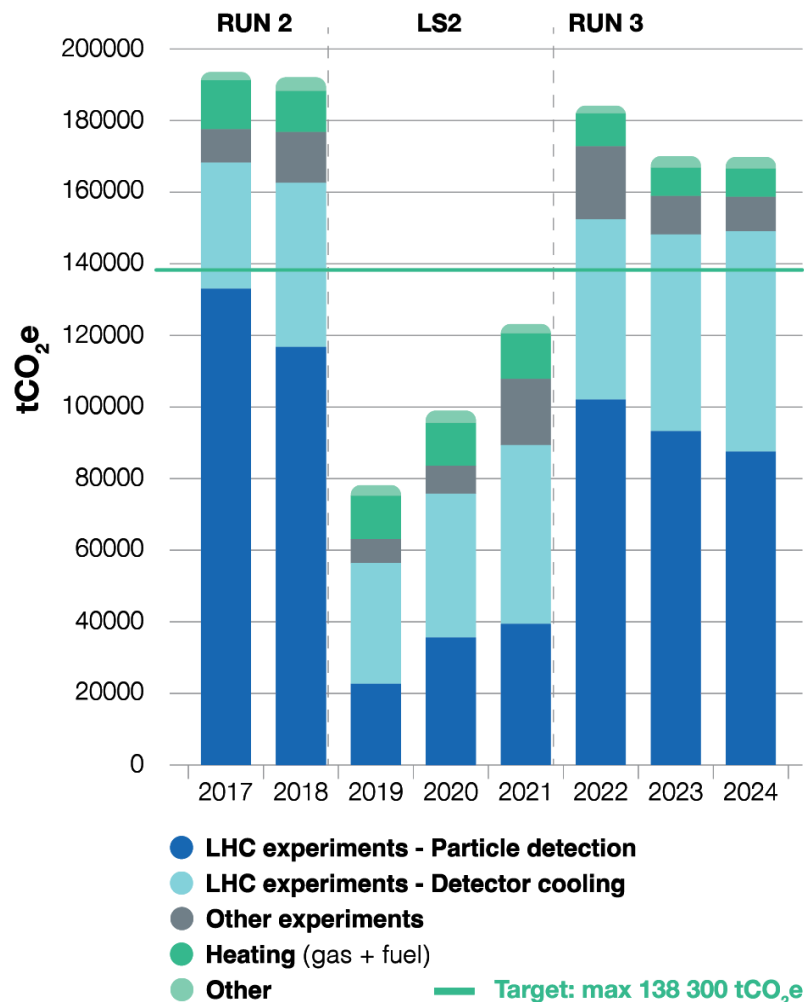
- ICT: approximately 4% of global emissions
- Aviation: approximately 2%

Global greenhouse gas emissions by sector
This is shown for the year 2016 – global greenhouse gas emissions were 49.4 billion tonnes CO₂eq.

Our World
in Data



CERN case study (electricity, scoped emissions, water)

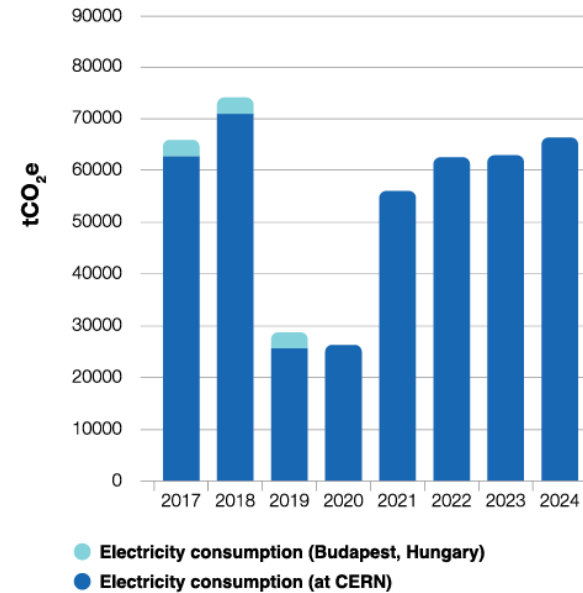


- CERN facility operates at 200 MW peak, performs winter shutdown to destress the grid.
- [1.3 TWh a year](#), < 1% of cloud.
- Scope 1 emissions (direct):
 - Includes:
 - Mainly synthetic gases and refrigerant production for particle detection, and cooling
 - Also includes building air conditioning/heating and other misc.
 - [~170 ktCO₂-eq in direct emissions as of 2024.](#)

CERN Scope 2 + 3 emissions

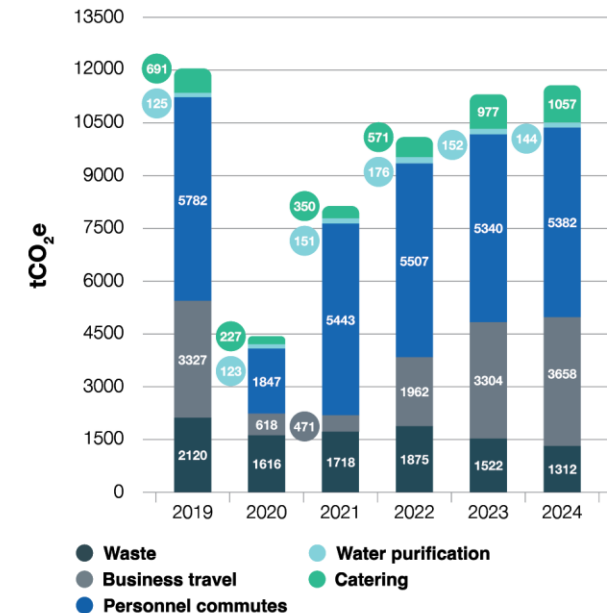


- Scope 2 (operational): CERN obtains mostly nuclear energy from EDF. Using emissions factors from the ADEME Base Empreinte, they report 67 ktCO₂-eq for 2024
- Scope 3 (upstream and downstream)
 - Includes:
 - waste treatment
 - water purification
 - travel
 - catering
 - Comes to about 11.5 ktCO₂-eq w/o procurement, but civil/mechanical engineering, electronics, and other scientific equipment adds up to ~100ktCO₂-eq
 - Important to considering manufacturing of required equipment and site-maintenance!



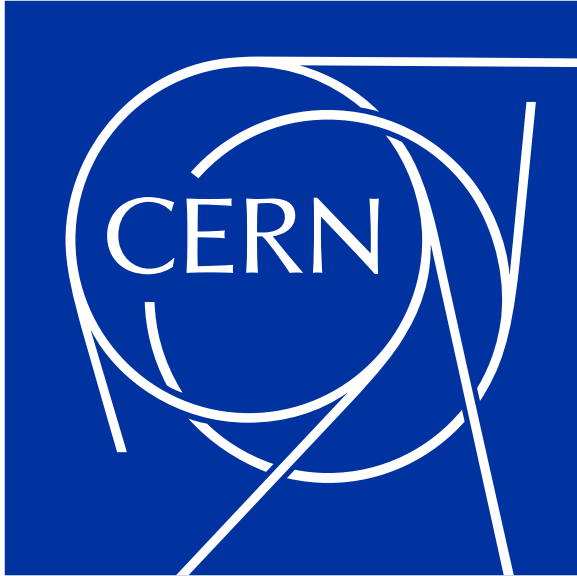
<- Scope 2

Scope 3 ->





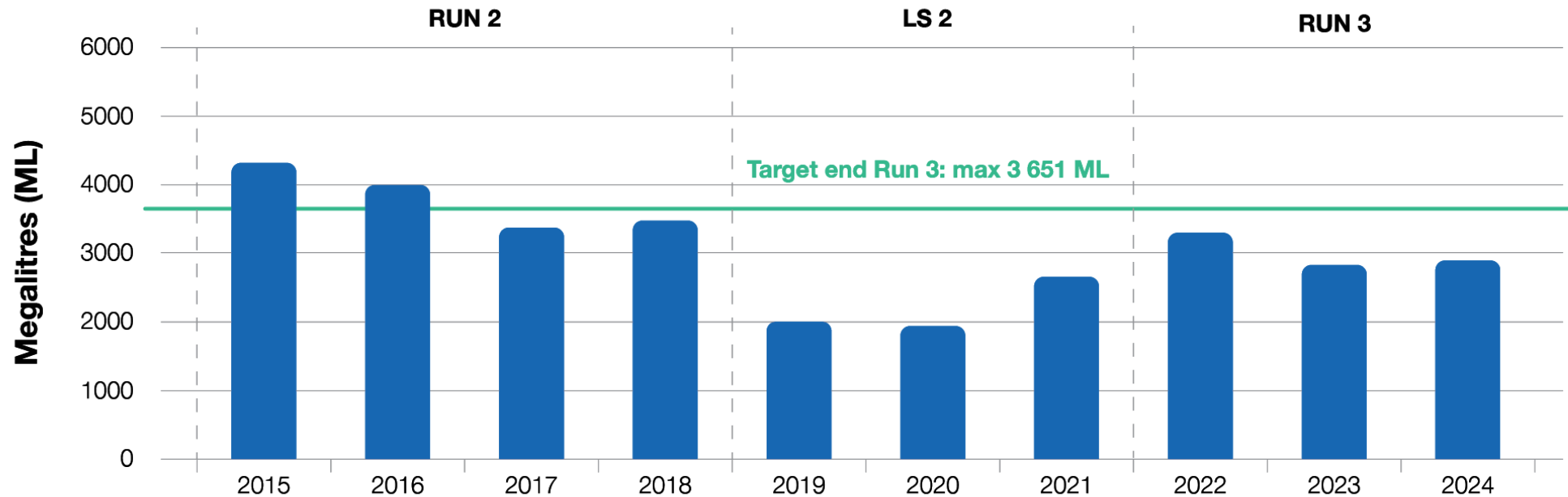
Total: 350ktCO₂-eq for 2024



Equiv. To 175,000 cars on the road annually



CERN uses about 2.9 B litres of water or 25k times the average UK household annual consumption from Lac Léman annually (mainly for cooling)



Now we have context, let's bring it back to the topic at hand



The largest AI data center campuses will soon be a fifth the size of Manhattan

EPOCH AI



Source: Contains modified Copernicus Sentinel data (2018)

Data center site boundaries are approximate.

CC-BY

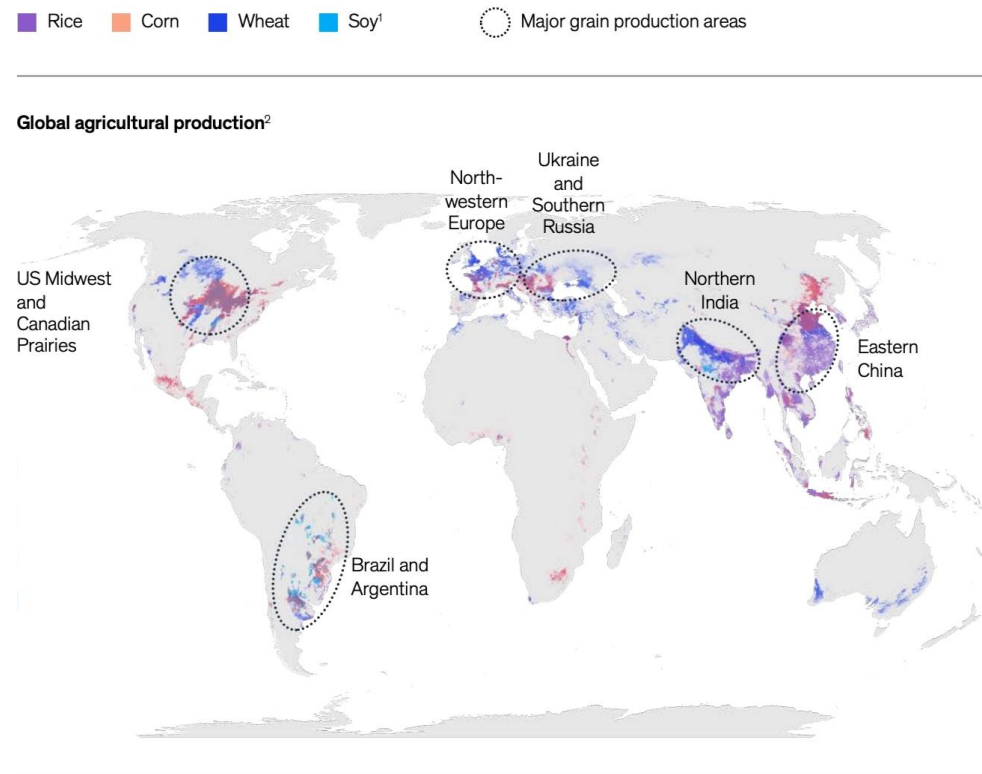
epoch.ai

- 30% of cloud spend is wasted due to over-provisioning:
 - Equivalent to 50 million UK households worth of annual electricity consumption, wasted.
 - Equivalent to 90% of the UK total annual car travel emissions. Wasted.

Allow me to indulge in a little climate fear-mongering



- 80% of the world population relies on maize, rice and wheat as primary food staples
- Specific regions of the world produce > 50% of calories consumed globally (1/4 of total cropland)



Allow me to indulge in a little climate fear-mongering



- There will be an [18% chance](#) of a “multiple bread basket failure” happening between 2030-2040 at current warming rates, up from 10% chance from 2020-2030, projected to increase to 34% between 2040-2050. That is if our global carbon emissions stop increasing soon - this could be much worse if we exceed 2°C.
- This would cause:
 - extreme price spikes
 - export restrictions
 - widespread malnourishment
 - mass migration
 - Potential supply chain breakdown
 - Resource wars?



Oh dear



That sounds pretty bad, but the digital sector can do their part to mitigate this scenario.

- Companies can implement frameworks to track IT usage via tagging and tracking spend to hold certain users and teams accountable.
- Aim to drive reduction in compute, storage and networking waste

Practical examples



- Downsizing of over-provisioned virtual machines and enforcement of auto-shutdown procedures outside of working hours
- Deleting unused stored data – [41% of data stored is never accessed again](#) according to industry reports
- Relying on batch-processing
 - Scheduling and processing workloads in parallel to saturate server compute and increase efficiency per workload and reduce overall emissions
- Use 'spot' instances where possible
 - Spare, idle compute capacity that were wasting power anyway for a much cheaper price, but you can be booted with a 2 minute warning



This is not straightforward

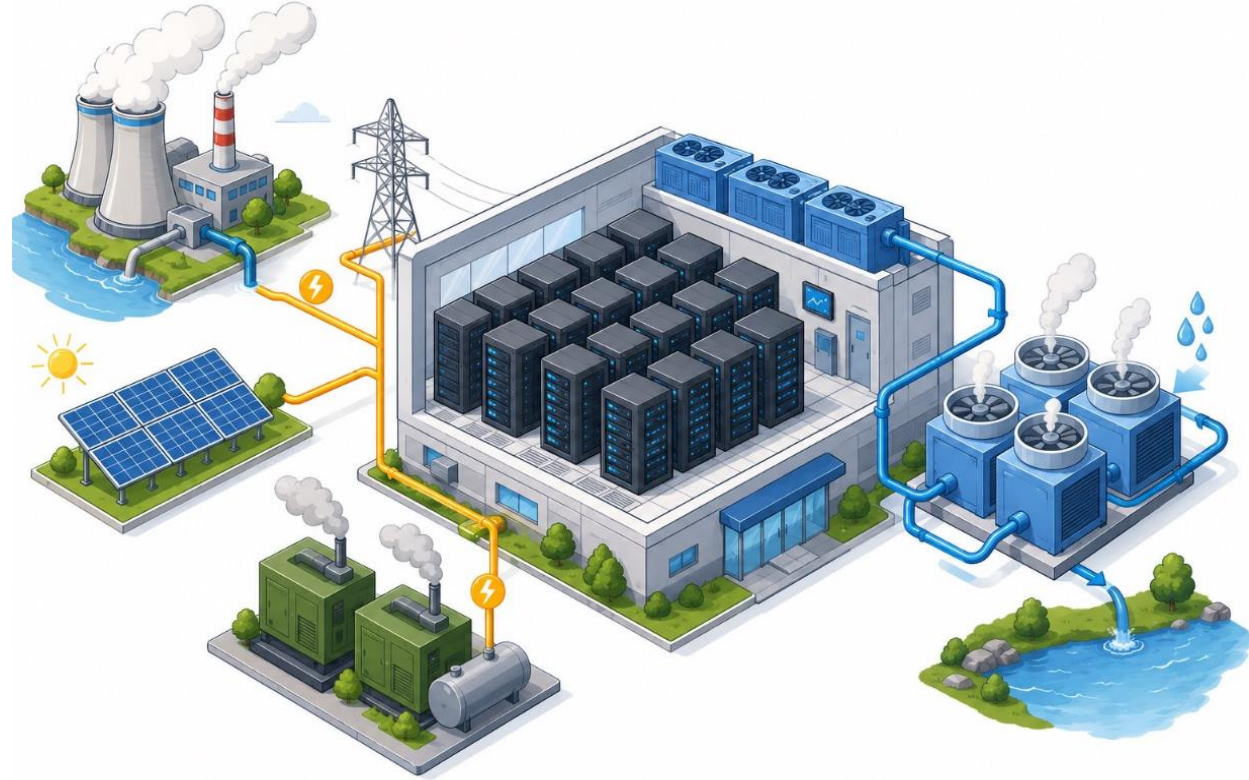


- All of these solutions require mindful engagement from employees in lots of different teams:
 - C-suite - drive an anti-waste culture, build initiatives
 - Engineering and cloud architects - taking a step back when building and deploying products
 - Finance and DevOps - tagging and tracking spend and communicating insights to other teams
 - Product - designing with frugality and efficiency in mind
 - Procurement - negotiate contracts only for what is needed
- This is hard to do!
 - Most employees don't care about how much they are spending - often none of their KPIs are to reduce spend.
- Need to try a multi-pronged approach to driving engagement.

Our dataset can be one of those prongs. What is in it?



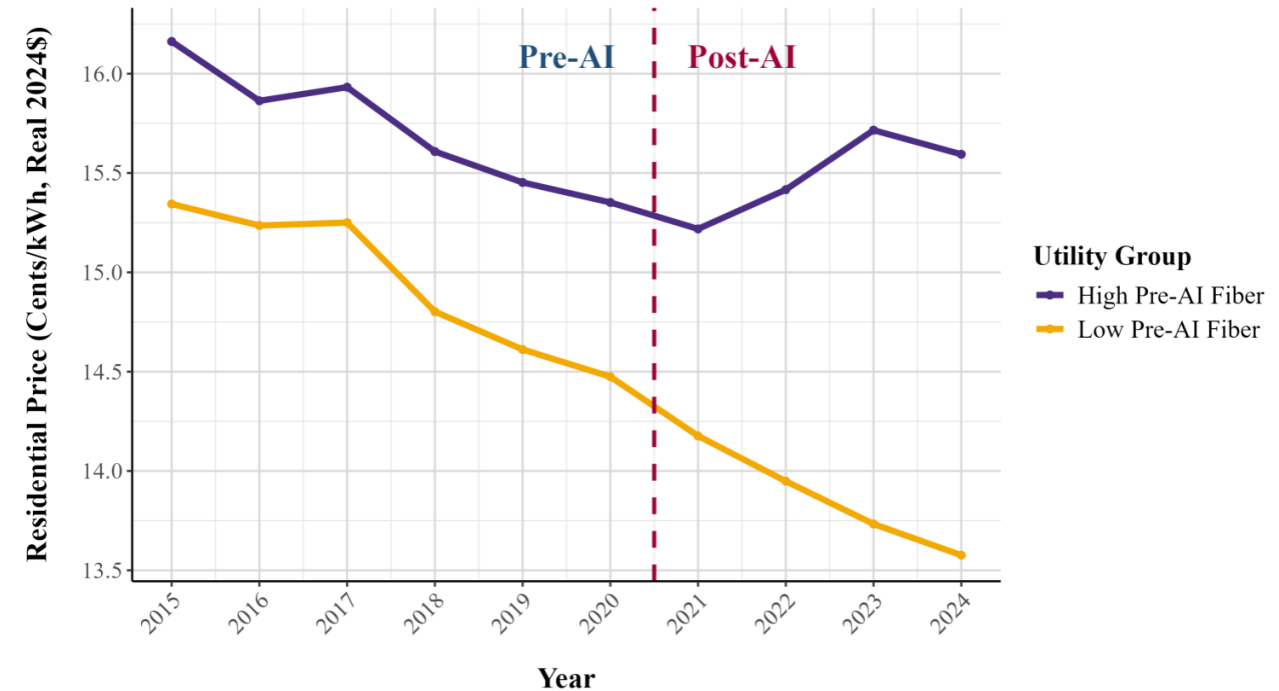
- Energy consumption - comes from the operation of hardware inside the hyperscaler. There are losses involved with the power efficiency of the data centre (cooling loops, transformer losses, lighting, monitoring etc.), which Greenpixie retrieves from publicly available data.
- Emissions - comes from onsite power generation, grid energy production, hardware manufacturing and transport, data centre constructions, and decommissioning.
- Water - used for data centre cooling, and power plant energy production.
- We estimate all of these as a function of energy consumed per business unit (e.g., a server allocated for one hour). Energy consumed is the fundamental unit we work from.



Data centre strain on local grids



- Energy consumption of data centres is only a small % of global consumption. However, data centres can cause local strain on grids requiring increased electricity importing and drives up local business and household rates.
- For example, in Virginia, USA, AKA “data centre valley”, [more than 25% of the grid power supply goes to data centres.](#)

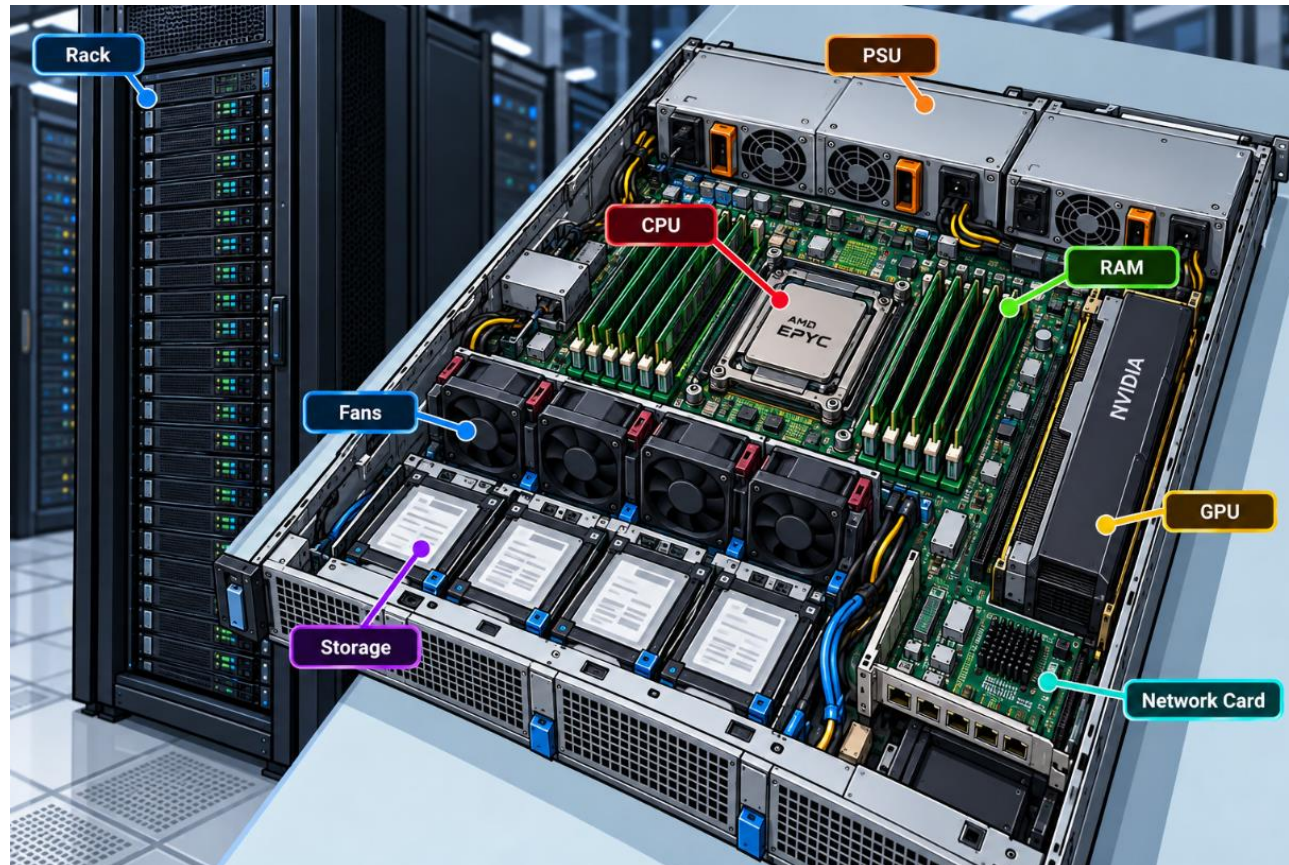


The [reversal of electricity price trends](#) in areas of the US with high-end fiber-optics infrastructure after the AI boom ^^

Estimating hardware energy consumption in datacentres



- Cloud servers look something like this. The power draw from each component is factored into the total kWh consumed by a service for a given business unit.



Estimating hardware energy consumption in datacentres



- For each compute instance across Amazon, Google and Microsoft cloud services, a hardware breakdown can be found in some capacity in their documentation.
Other services require a bit more detective work.

M7a

Instance Size	vCPU	Memory (GiB)	Instance Storage (GB)	Network Bandwidth (Gbps)***	EBS Bandwidth (Gbps)
m7a.medium	1	4	EBS-Only	Up to 12.5	Up to 10
m7a.large	2	8	EBS-Only	Up to 12.5	Up to 10
m7a.xlarge	4	16	EBS-Only	Up to 12.5	Up to 10
m7a.2xlarge	8	32	EBS-Only	Up to 12.5	Up to 10
m7a.4xlarge	16	64	EBS-Only	Up to 12.5	Up to 10
m7a.8xlarge	32	128	EBS-Only	12.5	10

[Amazon EC2 M7a instances](#), powered by 4th Generation AMD EPYC processors, deliver up to 50% higher performance compared to M6a instances.

Features:

- Up to 3.7 GHz 4th generation AMD EPYC processors (AMD EPYC 9R14)
- Up to 50 Gbps of networking bandwidth
- Up to 40 Gbps of bandwidth to the [Amazon Elastic Block Store](#) (Amazon EBS)
- Instance sizes with up to 192 vCPUs and 768 GiB of memory
- SAP-certified
- Built on the [AWS Nitro System](#), a combination of dedicated hardware and lightweight hypervisor
- Support for always-on memory encryption using AMD secure memory encryption (SME)
- Support for new processor capabilities such as AVX3-512, VNNI, and bfloat16.

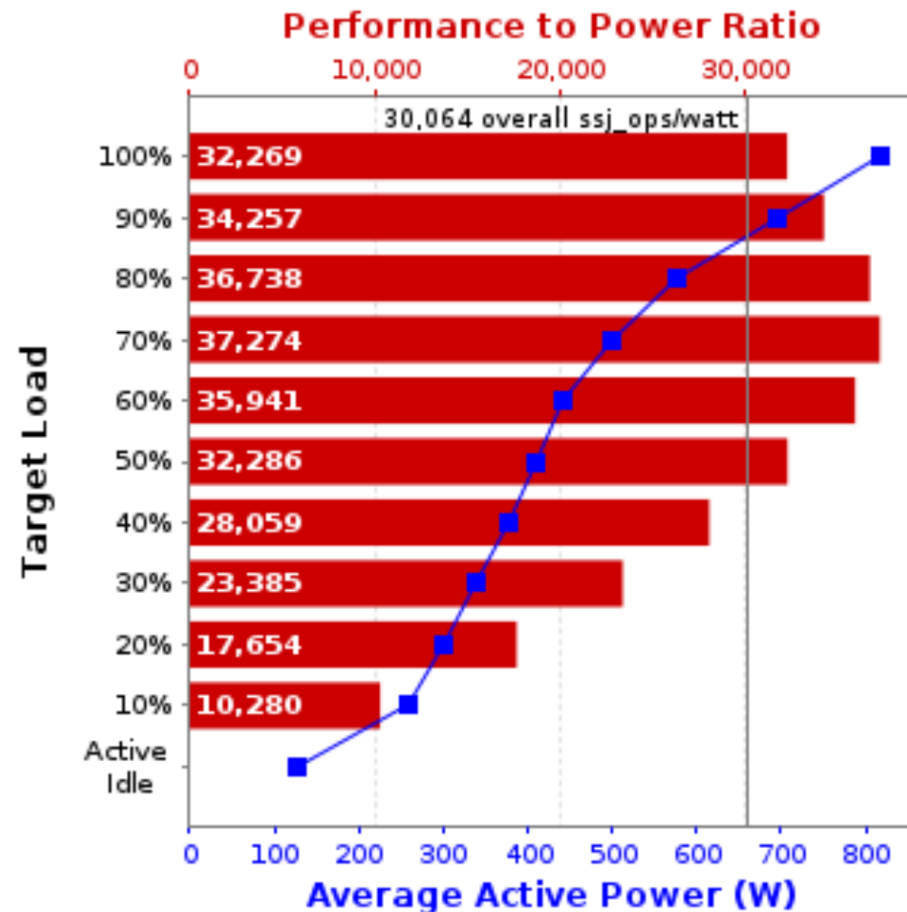
Use cases

Applications that benefit from high performance and high throughput such as financial applications, application servers, simulation modeling, gaming, mid-size data stores, application development environments, and caching fleets.

Estimating hardware energy consumption in datacentres



- We can then find power-utilisation curves from APIs (TechPowerUp) and public online benchmarks (SPEC) of each CPU/GPU/RAM model to estimate expected power draw per component depending on the type of service

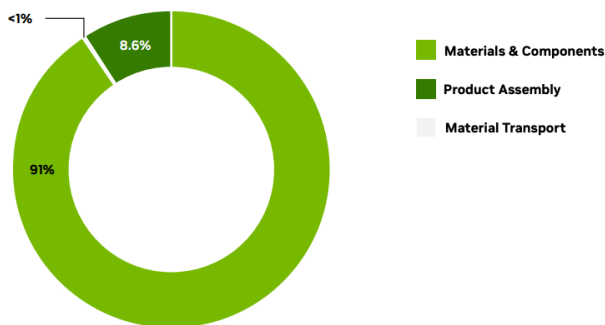


For the remaining components, in some cases proxies/assumptions must be made for the power draw of SSD/NIC/router/switch/fans/PSU/rack typically used in hyperscaler facilities, as this information is not always publicly available

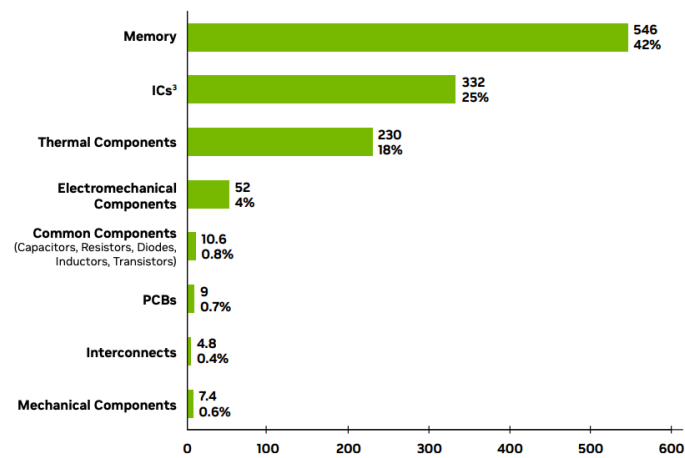
Estimating hardware manufacturing emissions



% Breakdown by Life Cycle Stage



HGX H100 PCF: Material Breakdown by Component Type (Kg CO2e/Unit)



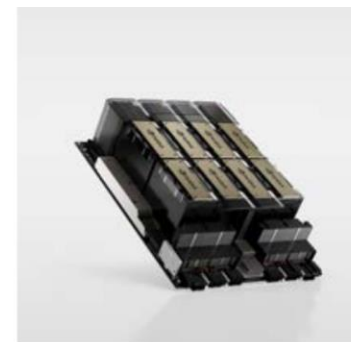
Disclaimer: The information in this document is solely for informational purposes and provides a general overview of potential emissions associated with the NVIDIA HGX H100 baseboard. NVIDIA's estimates in this document are based on currently available data. The information may contain estimates and certain data points and estimation methodologies are subject to change.
¹ ICs: Includes GPUs, network processors, and other integrated circuits.

- For hardware manufacturing emissions (scope 3), we research lifecycle assessment reports from manufacturers and academics to obtain the total embodied emissions associated with the production, transport and decommissioning of a certain piece of silicon
- Then we multiply that by the ratio of the time period that the component was used for, over the total lifespan of the component, for fair attribution.

Datasheet



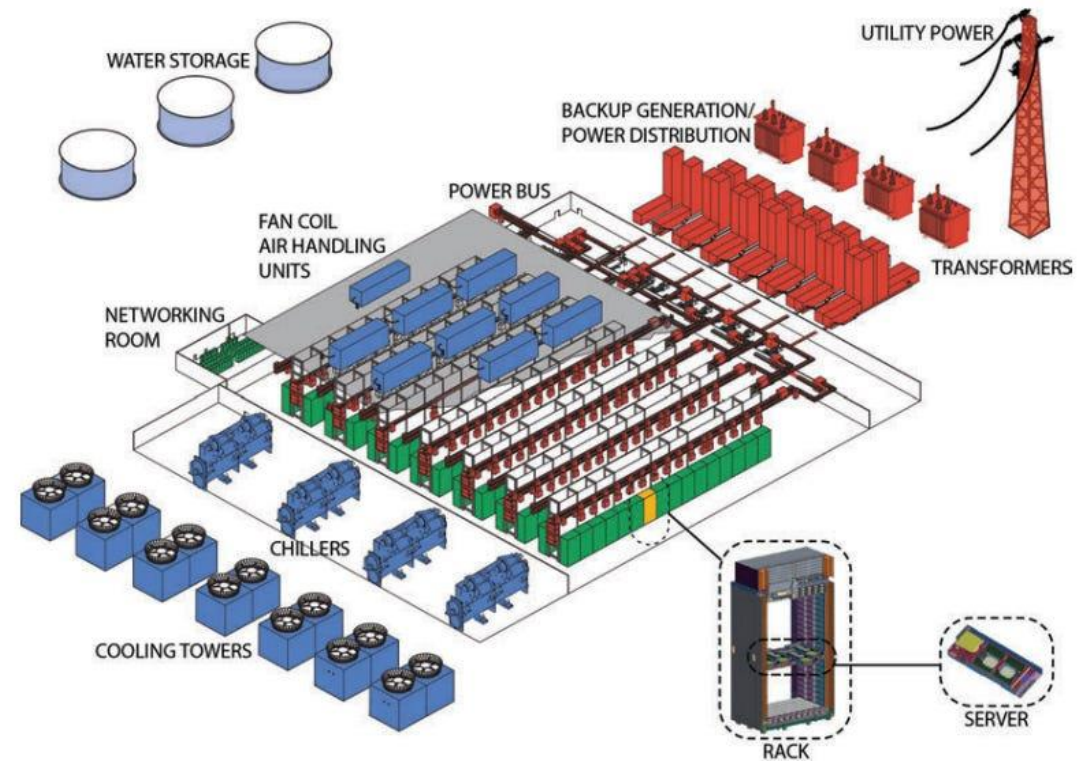
Product Carbon Footprint (PCF) Summary for NVIDIA HGX H100



Estimating manufacturing emissions



- We also account for the scope 3 emissions associated with building the data centre shell + miscellaneous equipment (sans servers) by looking at candidate data centre building reports and estimating the amortised amount of CO₂-eq per kWh consumed for a data centre of a given MWh capacity.



Estimating fuel mix and carbon intensity of each kWh used



Sample of fuel mix in London on 2026-09-15 at 03:00 UTC,
which averages to 116.7 g CO₂-eq per kWh

Fuel	Share of generation	Factor (g CO ₂ e per kWh)
gas	16.6%	490
biomass	11.7%	230
wind	56.7%	11.5
nuclear	14.8%	12
hydro	0.2%	24

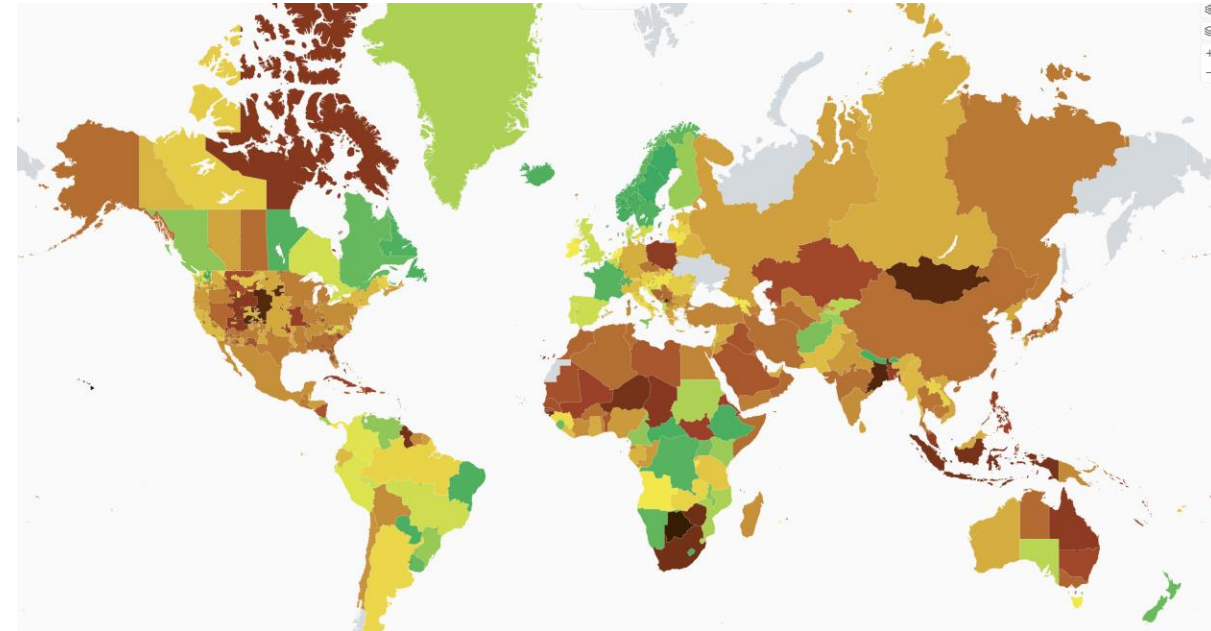
- To derive operational emissions we need to know which fuel types are used to produce energy during the hour/day/month the cloud service was used, and the amount of CO₂-eq emissions associated with energy produced by a given fuel in a given country.
- We perform a weighted average of the carbon intensity of each fuel type, the weights being derived from the total power produced by each present fuel type in a given time period for a certain region.
- We get this data from public APIs and web pages, for example for half-hourly UK data we use NESO (UK grid operator)

Estimating fuel mix and carbon intensity of each kWh used



- Renewables still have non-zero factors due to the amortised emissions of material mining, manufacturing, distribution and disposal.
- We account also for the electricity lost during transmission between the power plant and the data centre, which makes the carbon intensity higher since it takes more electricity + therefore carbon to power the data centre. This is not including the energy that doesn't make it to the data centre hardware due to facility losses.

Operational emissions = energy (kWh) × weighted average carbon intensity (g CO₂e per kWh)



Carbon intensity of electricity around the world. Green is low, dark brown is high. Source: Electricity Maps (app.electricitymaps.com).

News

- A lot of hype about data centre water consumption, among other issues like:
 - [Air pollution due to backup diesel generators](#)
 - [Noise pollution from the cooling system and generators](#)
 - [Heat islands](#)
 - [Not creating the job opportunities initially promised](#)

The endless data centre whine threatening to blight Britain's peaceful countryside

Residents living near giant AI facilities report invasive levels of noise

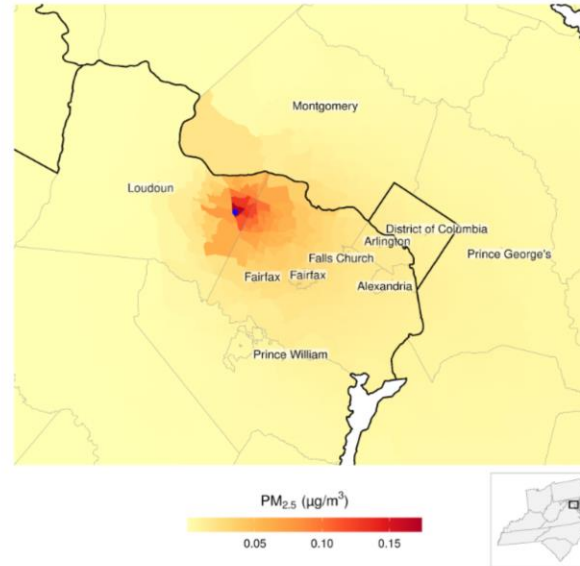


Figure 1 (above). Estimated regional distribution of annual-average $PM_{2.5}$ increases attributable to the Vantage facility.

The A.I. Data Center Backlash Has Reached a Quiet Corner of Rural England

Plans to build one of Europe's largest data centers in Devon have been met by strong opposition from residents, as environmental concerns grow.

Listen · 7:55 min

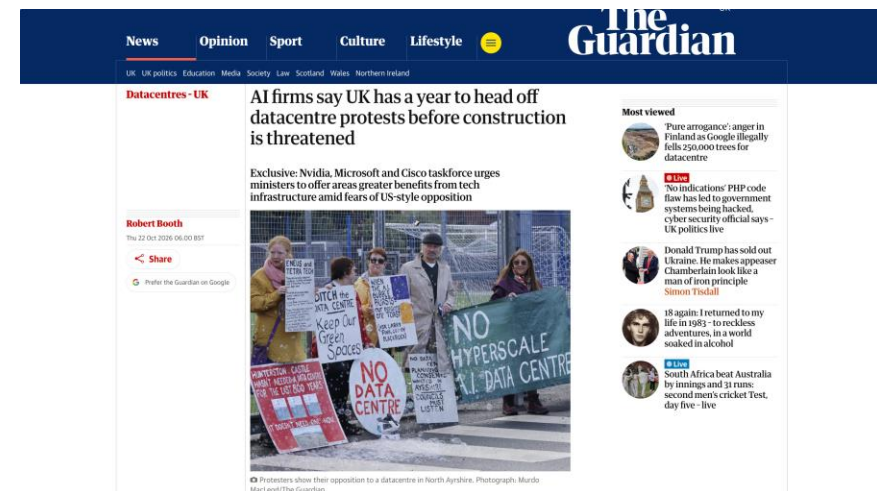
Share full article



53

A backlash over data centres is another threat to the AI juggernaut

16 September 2026 | 264 Comments



Water usage



- People say that water use by data centres is low in comparison to agriculture, and even golf courses.
 - US golf courses use ~550 billion gallons per year (usually non-potable and raw)
 - Data centres directly consume ~17 billions gallons per year (usually treated municipal water). However, this could be estimated to be much higher if scoped to power generation water consumption.
 - It is worth noting again that data centres can strain limited local water reserves. The locations that data centres are being constructed within are less ideal than those built 15-20 years ago at the advent of cloud, since all of the good spots with abundant water + power resources are already taken!



VS.



Water usage



- [It is all a trade-off](#). Dry cooling requires exorbitant quantities of peak power in the summer months, which can be mitigated by evaporative cooling but significantly spikes local water consumption.
- For estimating cooling water consumption, Amazon and Microsoft publish figures for each cloud region (us-east1, eu-west1 etc.) outlining the litres of cooling water evaporated per kWh of electricity consumed by the data centre hardware. Google published one equivalent global average number in a 2024 environmental report.
- For calculating water consumption from grid-purchased electricity (including imported electricity), The World Resources institute [wrote a report](#) providing litres per kWh coefficients for each country. Interestingly, heavily hydro-powered grids use lots of water due to reservoir evaporation.

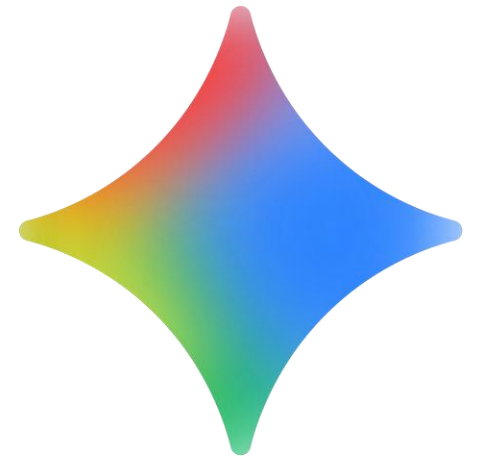
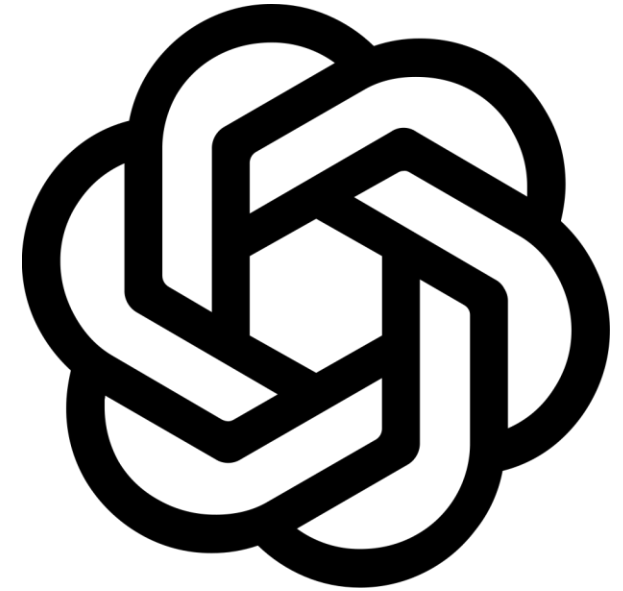
Total water consumed per kWh of data centre electricity used =

(hardware electricity consumed x the data centre cooling water factors) + (the electricity consumed at the facility level including efficiency losses x the grid-level water factors).

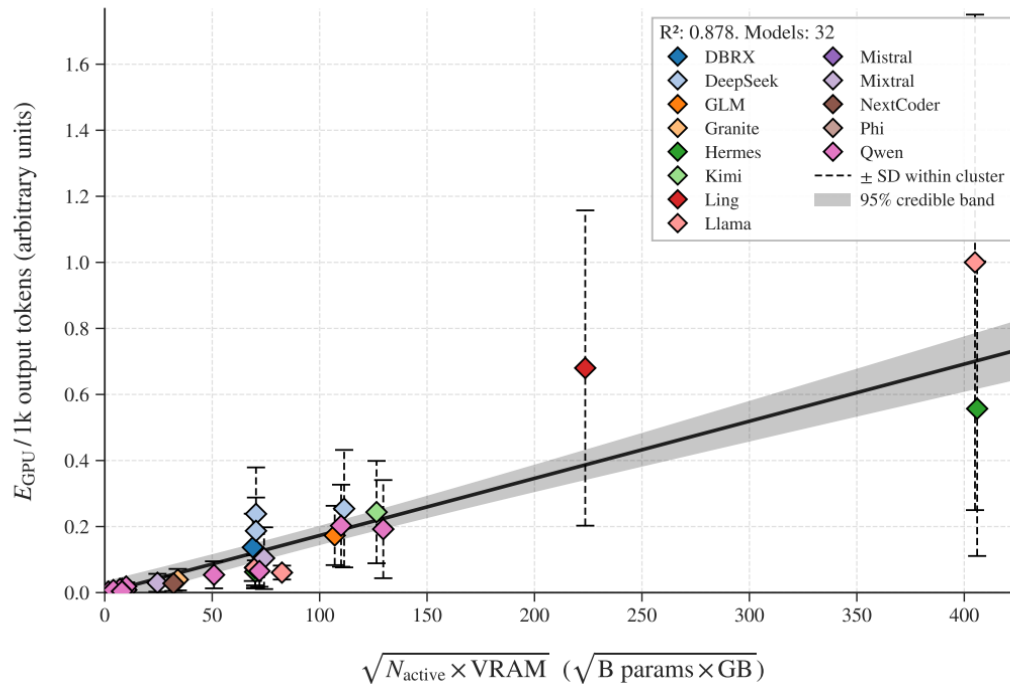
LLMs



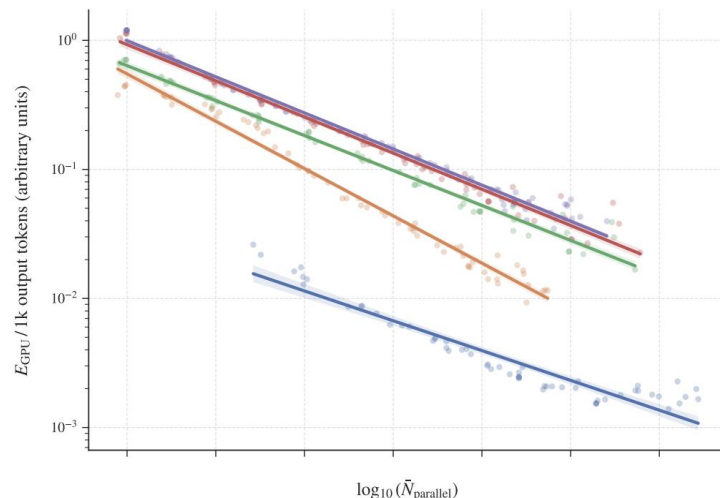
- The environmental impact of GenAI is harder to estimate because many providers don't want to disclose intellectual property to their competitors.
- They also know a lot of what they are doing to build the infrastructure for, and train AI is unethical (and sometimes illegal). Better to keep it all under wraps!



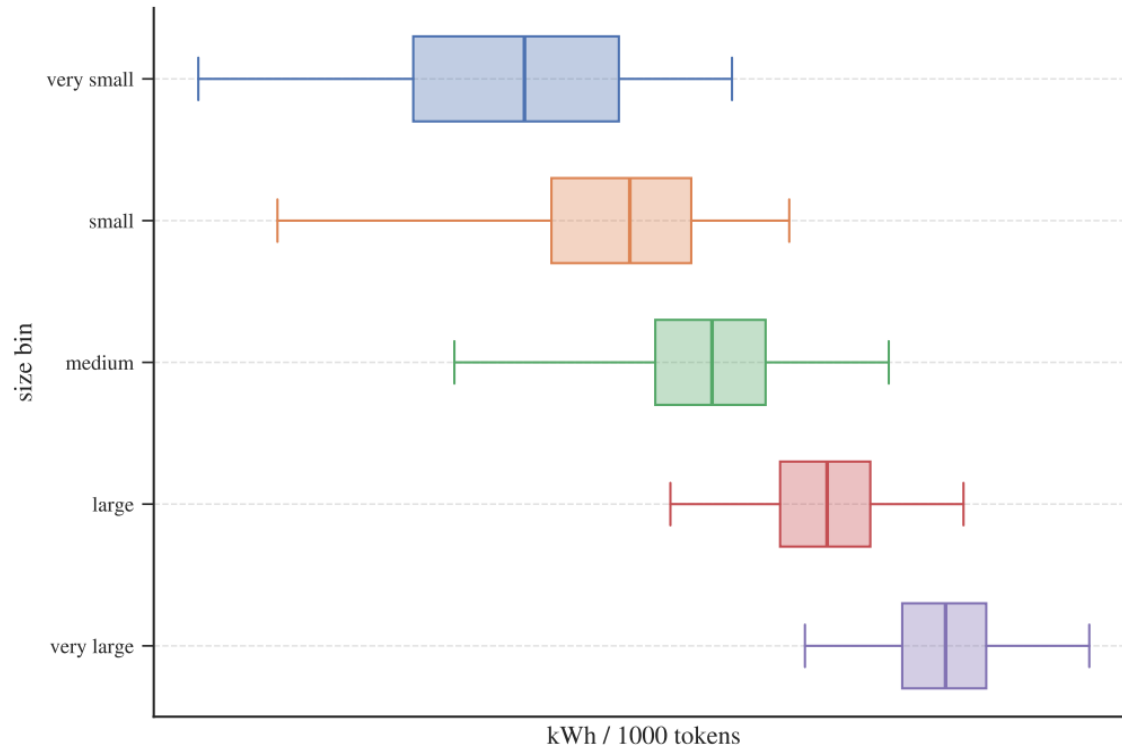
Measuring the impact of AI empirically



- What we can do is understand how the architecture of frontier open weights GenAI models (e.g., DeepSeek, Mistral, GLM, Kimi) affects energy consumption, depending on hardware, workload, cache-use, context length etc.
- I did this by renting some GPU machines and running inference benchmarks for a number of models earlier this year, and fitting a regression model to the resulting data.



Measuring the impact of AI empirically



- Then we can map these numbers to black-box proprietary LLMs (ChatGPT, Grok, Gemini, Claude) using public performance metrics/API token throughput/release date proxies, and guesses based on nomenclature.
- We make assumptions on weight quantisation, deployment hardware, number of API requests processed in parallel, LLM training energy consumption, and LLM architecture.
- We estimate the uncertainty on these assumptions by benchmarking and modelling all of these different configurations (FP4 vs. FP8, H100 vs. B200 GPU, low batching vs. high batching etc.) empirically and then randomly sampling the outputs of the models using a uniform prior.
- We take the IQR of the resulting distribution to form an upper and lower bound.

Measuring the impact of AI empirically



Greenpixie's AI Token Methodology

Assessing the Energy, Water and CO₂-eq Impact of AI Tokens for Open and Closed Weight Models

Joshua Horswill, Ross Hunter, Matt Clifford, James Hall

Greenpixie Ltd., London, UK.



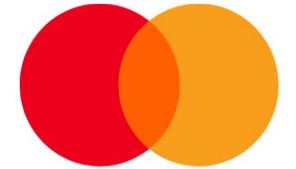
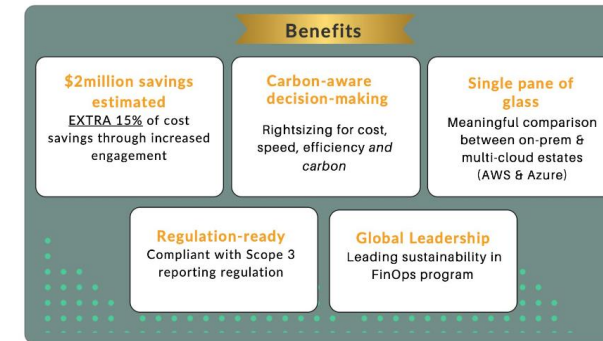
Some customer success stories



- Mastercard saved 1% of CO2-eq emissions despite 13% of company growth, and \$2M in cloud spend using our data to drive engagement. Similar success stories from ABB, and ABN AMRO
- A global manufacturer created an org-wide dashboard to name and shame teams and VP's responsible for excessive cloud carbon emissions and encourage financially responsible provisioning.
- A world bank used our data to estimate the infrastructure and power needed to shift their AI fleet from cloud to private on-prem deployments.
- An energy supplier wanted to encourage AI adoption from their employee base, as they were receiving bottom-up pressure regarding the ethics and environmental impact. They used our AI methodology to estimate their total AI emissions across all platforms to show that they were negligible in comparison to their scope 1 emissions, and that they were taking sustainability seriously.



How adding Sustainability to FinOps unlocks millions in cost savings



"Partnering with Greenpixie was a remarkable experience. Their seamless integration of existing tools and cutting-edge technology for cloud sustainability data set a high standard in the industry."



VIK SALUJA
Director of FinOps, TBM, and Sustainability
Mastercard

1% emissions reduction against 13% corporate growth

So, what have we covered?



- CERN's environmental footprint, and the impact of the digital sector
- The potential consequences of not being carbon-conscious
- How companies keep within their cloud budget, and the problems they face in the process + how sustainability data can help with that.
- Measuring sustainability metrics in cloud and for AI
- The impact that the data has in the corporate world.



If you want to find out more about cloud sustainability, look at Greenpixie's [free courses](#), and check out their [website](#).



Thank you for listening!



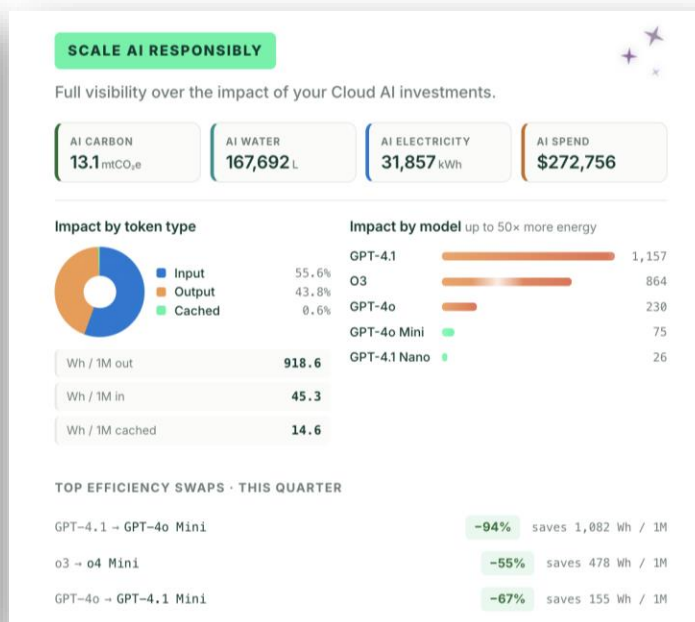
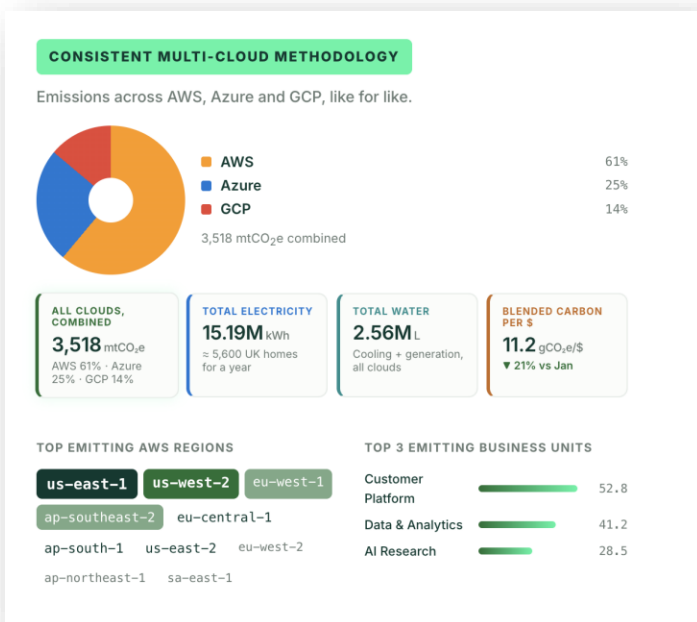
Appendix

Greenpixie sells a maintained coefficient database



11 million+

individually priced cloud items carry an energy, carbon and water coefficient



200+ AI models
alongside AWS,
Azure and Google
Cloud

ISO 14064-verified
methodology,
published in full

Greenpixie vs Native Tools

Required for effective FinOps & decarbonisation

Cloud Provider	Scope 1	Scope 2- location based <i>(Actual emissions)</i>	Scope 3	ISO-14064 verified	Water Metrics	Electricity Metrics	Enables AI impact analysis and tokenomics	Covers all services and usage types	Preserves all business tagging	Multi-cloud normalized methodology
		No transparency in which grid intensities used	All services that don't use dedicated racks rely on cost proxies		Uses same blended spend based allocation model as carbon	No electricity metrics shared		Only covers 3 services and an 'Other' category		
			Uses unclear attribution models and spend proxies	Not updated since 2021 and it only verifies allocation methodology	No water metrics shared	No electricity metrics shared		Do not publish a list of supported services		
		Claims to use electricity maps hourly data	Uses total electricity to allocate scope 3 contribution		No water metrics shared	No electricity metrics shared	Provides one number for all of Vertex with no breakdown	Covers 80 services		
		Up to hourly grid intensities for every cloud region	Including raw material extraction, manufacturing and end of life based on functional units	As of March 2024	Resource level water based exclusively on activity	The only provider including this	Per GPU, model, and token type metrics for almost 300 models	Covers 100% of services across all three cloud providers	Appends additional metrics onto existing usage reports	Azure, AWS & GCP supported



Moving to Graviton cuts energy, carbon and water more than cost

One AWS instance moved from x86 to Graviton and run for a year, 8,760 hours, in us-east-1

48%

less energy and water

by replacing one m5.xlarge with an m6g.xlarge, which also cuts carbon by 43% and cost by 20%

Before After

Instance move price per hour	Cost USD per year	Energy kWh per year	Carbon kgCO ₂ e per year	Water litres per year
m5.xlarge to m6g.xlarge \$0.192 to \$0.154	-20% 1,682 to 1,349	-48% 141.1 to 73.7	-43% 64.87 to 36.92	-48% 351 to 183
m6i.xlarge to m7g.xlarge \$0.192 to \$0.1632	-15% 1,682 to 1,430	-29% 127.5 to 90.5	-26% 58.22 to 43.11	-29% 317 to 225
c6i.2xlarge to c7g.2xlarge \$0.34 to \$0.29	-15% 2,978 to 2,540	-47% 223.4 to 117.8	-45% 101.35 to 56.02	-47% 556 to 293
r5.xlarge to r7g.xlarge \$0.252 to \$0.2142	-15% 2,208 to 1,876	-25% 204.3 to 153.7	-22% 94.34 to 73.31	-25% 509 to 382

Savings from rerouting models to models with similar intelligence

What it means	Intelligence change	Energy saved (kWh)	Cost saved
Swap Claude Opus 5 at high reasoning for GPT-6 Astra at low reasoning.	-4.6%	-2,461 (25.0%)	-\$39,551 (17.8%)
Swap Claude Fable 5.1 at high reasoning for GPT-6 Astra at medium reasoning.	-2.9%	-1,455 (14.8%)	-\$33,043 (14.9%)
Swap Claude Opus 5 at high reasoning for GPT-5.6 Sol at max reasoning.	-2.3%	-1,494 (15.2%)	-\$22,965 (10.3%)
Swap Claude Fable 5.1 at high reasoning for Claude Opus 5 at high reasoning.	-5.9%	-1,373 (14.0%)	-\$4,183 (1.9%)
Swap Claude Opus 5 at high reasoning for Claude Sonnet 5 at max reasoning.	-20.3%	-1,424 (14.5%)	+\$20,981 (9.4%)

Notes

- Scores and token counts are Artificial Analysis Intelligence Index v4.3, read 17 September 2026. Energy uses the Greenpixie AI Explorer rate card at low effective batch size.
- Intelligence change is measured against the model being replaced. Opus 5 at high scores 48.2 and Fable 5.1 at high scores 51.2 on the Intelligence Index.
- Savings rows do not factor in increased retries from lower intelligence so actual savings may be lower.

*A lower Intelligence Index means a smaller share of tasks are completed correctly at the first attempt, so the difference shows up as retries and human correction rather than as outright failure. 1% lower in intelligence equates to 1-3% more retries