UZH IT & AI Emissions Analysis

Massimiliano Poletto

Last modified: 18.12.2024

1 Introduction

We are trying to answer two questions:

- Are there emissions savings in moving UZH's IT infrastructure to the cloud?
- What is the impact of increased AI use (and, in particular, of LLMs) on UZH's emissions?

To this end, we need some UZH-specific information:

1. The size of UZH's student and staff population and a rough breakdown of what kind of work they do (to estimate their use of IT and AI).
2. Information about UZH's current IT infrastructure (number of servers, classroom equipment, etc.) and the energy use (electricity bills) of any campus data centers.
3. An estimate of the cloud compute capacity required to host UZH's current IT requirements.

In addition, we need information about the possible location of cloud data centers and LLM deployments, the embodied emissions of different types of IT equipment, and the carbon intensity of electricity in different regions.

We describe our modeling approach and our assumptions and uncertainties regarding these factors.

2 Modeling approach

At a high level, we model IT emissions primarily bottom-up, based on device counts and estimates of their characteristics (embodied emissions, power utilization, and duty cycle).

The model can operate in two modes: deterministic, where each device type has a fixed set of characteristics based on the mean of values we found in the literature, and stochastic, where we pick device characteristics from truncated normal distributions with the mean and standard deviation of the literature values. Of course, normality is a strong and probably incorrect assumption, but it is a simple initial way to model the uncertainty in our estimates.

2.1 Campus emissions

We instantiate as many devices as specified in the UZH spreadsheet, plus enough laptops and phones to cover the rest of the university population.

For each device instance, we pick embodied emissions, power draw, and duty cycles either deterministically or by sampling from the appropriate distribution.

Combining power draw and duty cycle estimates with ground-truth data about the Swiss electricity mix, we estimate annual operational emissions and add them to the embodied emissions scaled by the device type's expected lifetime.

For on-campus data centers, for which presumably there are electricity bills or other measures of overall energy use, we would like to use that data to evaluate our model.

A possible future refinement will be to model the duty cycle as a function of time of day and incorporate finer-grained electricity mix data.

2.2 Cloud emissions

We model cloud emissions based on the assumptions and data described in Section 3.4. We suggest working with UZH's cloud vendor to obtain more accurate data.

2.3 LLM emissions

We model LLM emissions by sampling from the per-capita daily query distribution and computing the carbon intensity of each query based on the average European electricity mix. While there are many uncertainties about the level of LLM adoption, there appears to be reasonable consensus in the literature about the energy consumption of each query.

3 Data sources and unknowns

3.1 UZH population

We have access to a spreadsheet (`2024-11-13_Berechnung_THG_IKT_2024.xls`) with counts and emissions factors for various types of IT equipment at UZH. It reports approximately 10,000 laptops and monitors, 1,000 servers, 2,400 mobile phones, 150 classroom PCs, and a few hundred printers, copiers, and projectors.

By contrast, Zürich (2024) reports approximately 28,000 students and 7,300 staff at UZH. Evidently the spreadsheet only lists university assets, not personal devices.

In modeling overall university IT emissions as well as potential LLM use, we assume that 90% of the university population, 32,000 people, have a laptop and a mobile phone. We do not yet have any information about the breakdown of different functions / job roles, so we assume that IT use is uniform across the university. Both of these assumptions should be revisited.

3.2 Embodied emissions

For several device types, especially end-user devices such as phones and laptops, embodied emissions comprise a majority of lifetime emissions. However, the exact amount of these emissions varies widely between different sources. Our approach has been to find several studies and industry references, and then estimate means and standard deviations from those that seemed most comprehensive and reliable (for example, Dell's Product Carbon Footprint reports about its servers).

3.2.1 Laptops

Production & Disposal (kg CO_2e)	**Source**
121	Ecoinvent (2024)
104	Teehan and Kandlikar (2013) (low)
338	Teehan and Kandlikar (2013) (high)
244	rarecoil (2021)
202	*UN Digital Economy Report* (2024)

The Ecoinvent numbers appear to be outliers on the low end. The most comprehensive data comes from rarecoil (2021), which lists manufacturer information for over 90 popular devices. Mean supply chain emissions are 244 kg CO_2e with a standard deviation of 128 kg. We model laptop supply chain emissions using a truncated normal distribution with those characteristics.

3.2.2 Desktops

Production & Disposal (kg CO_2e)	**Source**
238	Ecoinvent (2024)
303	Teehan and Kandlikar (2013)
403	*UN Digital Economy Report* (2024)
289	Dell Technologies (2024)
277	Boavizta (2024)

Most estimates are in the range of 250 - 400 kg CO_2e. The most detailed data comes from Dell's Product Carbon Footprints (Dell Technologies (2024)), which they publish for each of their products. A sample of 15 desktops has mean supply chain emissions of 289 kg CO_2e with $\sigma = 80$ kg. We model desktop supply chain emissions using these statistics.

3.2.3 Servers

Production & Disposal (kg CO_2e)	**Source**
383	Teehan and Kandlikar (2013)
1252	Davy (2021) (Dell)
1912	Davy (2021) (EC2)
899	Boavizta (2024) (medium server)

Davy (2021) reports data for many Dell servers and estimates for different types of EC2 instances. The former have mean supply chain emissions of 1252 kg CO_2e with $\sigma = 330$ kg, the latter 1912 kg with $\sigma = 885$ kg. We model server supply chain emissions using the statistics corresponding to Dell servers.

3.2.4 Smartphones

Production & Disposal (kg CO_2e)	Source
68	Google (2023) (Pixel 8)
54	Apple (2021) (iPhone 13)
50	*UN Digital Economy Report* (2024)
50	Lövehagen et al. (2023)

We model phone supply chain emissions using mean 50 kg CO_2e and $\sigma = 10$ kg.

3.2.5 Other devices

We found less data for other types of devices. For now, we model them with averages based on a handful of sources.

Computer monitors 344 kg CO_2e (Teehan and Kandlikar (2013), Dell Technologies (2024))

Conference room displays 753 kg CO_2e (scaling monitor by $(40''/27'')^2$)

Printer/copier stations 1167 kg CO_2e (Ecoinvent (2024))

Network equipment We did not find reliable lifecycle assessments for routers and switches, and have no information about UZH network architecture. The vast majority ($80 - 95\%$) of lifecycle emissions of network gear stem from usage (Cisco Systems (2024), Jacob (2023)). For now, we model network equipment as a 5% overhead on servers (1 router or switch for 20 servers). (A typical data center rack contains 42 1U servers and one switch or router.)

3.3 Operational emissions

For on-premise equipment, we would like to model the Swiss electricity mix (see below) and, ideally, daily variations in usage and electricity carbon intensity. Generic manufacturer estimates of lifetime usage-related emissions are therefore unsuitable. There are surprisingly few sources of data on device power consumption (as opposed to CO_2 emissions). Based on a smattering of data sheets, primarily from Apple, Cisco, and Dell, we model power draw of different devices while under load using the following initial parameters:

Device	Mean (W)	σ (W)
Laptop	30	5
Desktop	100	20
Server	400	100
Phone	5	2
Monitor	50	10
Conference display	250	50
Printer/copier	1000	200

We assume a distribution of duty cycles and power draw when idle for each device. In the future, we may simulate power draw at different times of day to integrate with hourly electricity mix data, but errors in those assumptions may swamp the relatively small hourly variations in Swiss electricity carbon intensity.

3.4 Cloud data centers

The secretiveness of major cloud providers makes it difficult to estimate the embodied and operational emissions of cloud infrastructure. The best resource we have been able to find is the Datavizta API (Boavizta (2024)).

Lacking information about UZH's workloads, we model a hypothetical UZH cloud footprint as follows:

- One-for-one replacement of one campus server with one Azure D8S_v3 instance (2x Intel Xeon, 8 cores, 32 GB RAM, 2TB SSD) (Microsoft Corporation 2024).
- Deployment to Microsoft Azure's "Switzerland North" region .
- Estimation of annual emissions via Datavizta, assuming constant 50% load: 31 kg CO_2e for usage, 97 kg CO_2e for manufacturing.

3.5 Large language models

Energy consumption of LLMs is being studied extensively (Budennyy et al. (2022), Castano et al. (2023), De Vries (2023), Gowda et al. (2024), Harding et al. (2024), Heguerte et al. (2023), Luccioni et al. (2022), Luccioni and Herandez-Garcia (2023), Patterson et al. (2021), Rodriguez et al. (2024), Tripp et al. (2024)).

The best estimates are that use of large language model (LLM) like ChatGPT averages 3-4 Wh / request, or approximately 10x the energy of a traditional search engine query. This number appears to be decreasing rapidly as model hardware and software improve, even though overall energy use is increasing due to increased utilization. Nevertheless, for the model we assume that queries consume 3 Wh each. We assume that on average every user will issue 15 queries per day.

LLMs are distributed and the location that serves any particular request is unknown, so we use the average European electricity mix to compute carbon intensity.

The embodied emissions and training emissions of LLMs are difficult to estimate and attribute precisely. However, De Vries (2023) provides the following information about ChatGPT:

- Operational power consumption of 500 MWh / day.
- Approximately 1300 MWh consumed in training.
- Approximately 4000 servers.

Assuming a relatively clean grid at 200 g CO_2e / kWh (a low estimate), and that each server has embodied emissions of 4 t CO_2e (a high estimate), we obtain:

- 100 t CO_2e / day operational emissions.
- 260 t CO_2e training emissions.
- 16 kt CO_2e total embodied emissions.

After one year of operation, operational emissions are approximately 36 kt CO_2e, whereas training and embodied emissions are about 16 kt CO_2e, so operational emissions account for $\approx 70\%$ of the total. With less conservative assumptions (dirtier grid, lower embodied emissions), the proportion of operational emissions increases further.

As a result, when modeling overall LLM emissions, we multiply estimated operational emissions by 1.4 to account for training and embodied emissions. We acknowledge that this is, at best, a rough estimate based on just one LLM.

3.6 Electricity carbon intensity

The Swiss Federal Office for the Environment published an extensive report (Krebs and Frischknecht (2018)) on the environmental impact of electricity generation in Switzerland in 2018.

However, for the purposes of the model, we use data from Electricity Maps (2024). Specifically, we use the carbon emissions for electricity *consumption* (not generation) in Switzerland for 2023 at hourly granularity, then aggregate the data into two-hour blocks, and compute the mean of each two hour block over the course of the year. This gives us twelve data points that describe average carbon intensity variation over the course of the day.

4 Missing data

The following data would improve the accuracy of our model (in roughly decreasing order of utility):

- Campus data center energy use (e.g., at the granularity of monthly electricity bills).
- Current campus data center workloads (to better model transition to the cloud).
- Academic calendar (e.g., how many days per year is the university full vs operating at lower levels).
- Common or otherwise representative pieces of equipment (servers, routers, standard desktops, etc.).
- Rough breakdown of UZH's population by field / role.

5 Other considerations

5.1 Other IT components

For now, we do not attempt to model changes in network infrastructure or utilization due to move to the cloud or increased LLM use. See Jacob (2023) for an overview of network sustainability.

5.2 Potential AI benefits

We typically think of AI use as creating additional energy demands and GHG emissions. So far, this model does not include potential energy savings caused by AI, such as improvements to building energy management, campus logistics, etc.

Tomlinson et al. (2024) make the argument that tasks such as writing and illustrating can be done with fewer GHG emissions by AI than by humans. They compare the estimated emissions of an LLM performing the task to those of the average human, calculated by multiplying per-capita annual emissions by the time ($\approx$ 1h) required to write or draw. Using this approach, AI is orders of magnitude cheaper than humans, whether one uses per-capita emissions from the US (15 t CO_2e/y) or India (1.9 t). Of course, this methodology focuses on a specific creative act: the LLM does not reduce the human's overall annual emissions.

References

Apple (2021). *iPhone 13 Product Environmental Report.* URL: https://www.apple.com/environment/pdf/products/iphone/iPhone_13_PER_Sept2021.pdf.

Boavizta (2024). *Boavizta Datavizta API.* Methodology available at https://www.boavizta.org/en/blog/empreinte-de-la-fabrication-d-un-serveur. URL: https://dataviz.boavizta.org/.

Budennyy, S. et al. (2022). “Eco2AI: Carbon Emissions Tracking of Machine Learning Models as the First Step Towards Sustainable AI”. In: *arXiv preprint 2208.00406v2*. URL: https://arxiv.org/pdf/2208.00406.

Castano, J. et al. (2023). “Exploring the Carbon Footprint of Hugging Face’s ML Models: A Repository Mining Study”. In: *arXiv preprint 2305.11164v3*. URL: https://arxiv.org/pdf/2305.11164.

Cisco Systems (2024). *Cisco Product Sustainability*. URL: https://www.cisco.com/c/m/en_us/about/csr/esg-hub/environment/product-sustainability.html.

Davy, B. (2021). *AWS EC2 Carbon Footprint Dataset*. Teads Engineering. URL: https://medium.com/teads-engineering/building-an-aws-ec2-carbon-emissions-dataset-3f0fd76c98ac.

De Vries, A. (2023). “The Growing Energy Footprint of Artificial Intelligence”. In: *Joule* 7.10. URL: https://www.cell.com/joule/fulltext/S2542-4351(23)00365-3.

Dell Technologies (2024). *Product Carbon Footprints*. URL: https://www.dell.com/en-uk/dt/corporate/social-impact/advancing-sustainability/climate-action/product-carbon-footprints.htm.

Ecoinvent (2024). *Lifecycle Inventory Database*. URL: https://ecoinvent.org.

Electricity Maps (2024). *Electricity Maps Data Portal*. Swiss data available at https://www.electricitymaps.com/data-portal/switzerland. URL: https://www.electricitymaps.com/.

Google (2023). *Pixel 8 Environmental Report*. URL: https://sustainability.google/reports/pixel-8-pro-product-enviromental-report/.

Gowda, S. et al. (2024). “Watt for What: Rethinking Deep Learning’s Energy-Performance Relationship”. In: *arXiv preprint 2310.06522v2*. URL: https://arxiv.org/pdf/2310.06522.

Harding, A. et al. (2024). “Watts and Bots: The Energy Implications of AI Adoption”. In: *arXiv preprint 2409.06626*. URL: https://arxiv.org/pdf/2409.06626.

Heguerte, L. et al. (2023). “How to Estimate Carbon Footprint when Training Deep Learning Models? A Guide And Review”. In: *arXiv preprint 2306.08323v2*. URL: https://arxiv.org/pdf/2306.08323.

Jacob, R. (2023). *Sustainable Networking*. ETH Zürich. URL: https://nsg.ee.ethz.ch/files/public/slides/2023-12-19_AdvNet_SustainableNetworking_1.pdf.

Krebs, L. and R. Frischknecht (2018). *Umweltbilanz Strommixe Schweiz 2018*. URL: https://www.bafu.admin.ch/dam/bafu/de/dokumente/klima/fachinfo-daten/Umweltbilanz-Strommix-Schweiz-2018-v2.01.pdf.download.pdf/Umweltbilanz-Strommix-Schweiz-2018-v2.01.pdf.

Lövehagen, N. et al. (2023). “Assessing embodied carbon emissions of communication user devices by combining approaches”. In: *Renewable and Sustainable Energy Reviews* 183. URL: https://doi.org/10.1016/j.rser.2023.113422.

Luccioni, S. et al. (2022). “Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model”. In: *arXiv preprint 2211.02001*. URL: https://arxiv.org/pdf/2211.02001.

Luccioni, S. and A. Herandez-Garcia (2023). “Counting Carbon: A Survey of Factors Influencing the Emissions of Machine Learning”. In: *arXiv preprint 2302.08476v1*. URL: https://arxiv.org/pdf/2302.08476.

Microsoft Corporation (2024). *Azure Virtual Machine Series*. URL: https://azure.microsoft.com/en-us/pricing/details/virtual-machines/series/.

Patterson, D. et al. (2021). “Carbon Emissions and Large Neural Network Training”. In: *arXiv preprint 2104.10350*. URL: https://arxiv.org/pdf/2104.10350.

rarecoil (2021). “Laptop carbon footprints”. URL: https://github.com/rarecoil/laptop-co2e.

Rodriguez, C. et al. (2024). “Evaluating the Energy Consumption of Machine Learning: Systematic Literature Review and Experiments”. In: *arXiv preprint 2408.15128v1*. URL: https://arxiv.org/pdf/2408.15128.

Teehan, P. and M. Kandlikar (2013). “Comparing Embodied Greenhouse Gas Emissions of Modern Computing and Electronic Products”. In: *Environmental Science and Technology* 47.9. URL: https://pubs.acs.org/doi/10.1021/es303012r.

Tomlinson, B. et al. (2024). “The Carbon Emissions of Writing and Illustrating are Lower for AI than for Humans”. In: *Scientific Reports* 14.1. URL: https://doi.org/10.1038/s41598-024-54271-x.

Tripp, C. et al. (2024). “Measuring the Energy Consumption and Efficiency of Deep Neural Networks: An Empirical Analysis and Design Recommendations”. In: *arXiv preprint 2403.08151v1*. URL: https://arxiv.org/pdf/2403.08151.

UN Digital Economy Report (2024). Tech. rep. United Nations Conference on Trade and Development. URL: https://unctad.org/publication/digital-economy-report-2024.

Zürich, Universität (2024). “UZH in Zahlen”. In: URL: https://www.uzh.ch/de/explore/portrait/figures.html.