WIEVIEL CO2 EMITTIERT KÜNSTLICHE INTELLIGENZ?

Hendrik Weichel
Gian Luca Buono
Prof. Dr. Jörg Schäfer
Prof. Dr. Martin Simon

HPC CLUSTER DER HOCHSCHULE

Im Rahmen des Projekts KI-NACHWUCHS@FH hat die Frankfurt University of Applied Sciences im Jahr 2023 die Förderung für einen High-Performance Computing (HPC) Cluster erhalten.

Die aktuelle Omnipräsenz generativer KI wird u.a. durch massiv parallele Hardware, sog. Graphical Processing Units (GPUs), befeuert. Diese erlauben es, große Datenmengen schnell und effizient zu verarbeiten, was die Entwicklung von rechenzeitintensiven KI-Algorithmen ermöglicht.

Die Vorteile von Graphical Processing Units (GPUs) sind unter anderem:

- Massive Parallelverarbeitung
- Schnelle Speicherzugriffe
- Hohe Rechenleistung

Diese Infrastruktur ermöglicht den Forscherinnen und Forschern der Hochschule die Mitarbeit an zukunftsweisendne KI-Forschungsprojekten in verschiedenen Anwendungsdomänen wie etwa

- Große Sprachmodelle (LLMs)
- Industrielle Zeitreihenanalyse
- Finanzmarktmodelle
- Wettervorhersage

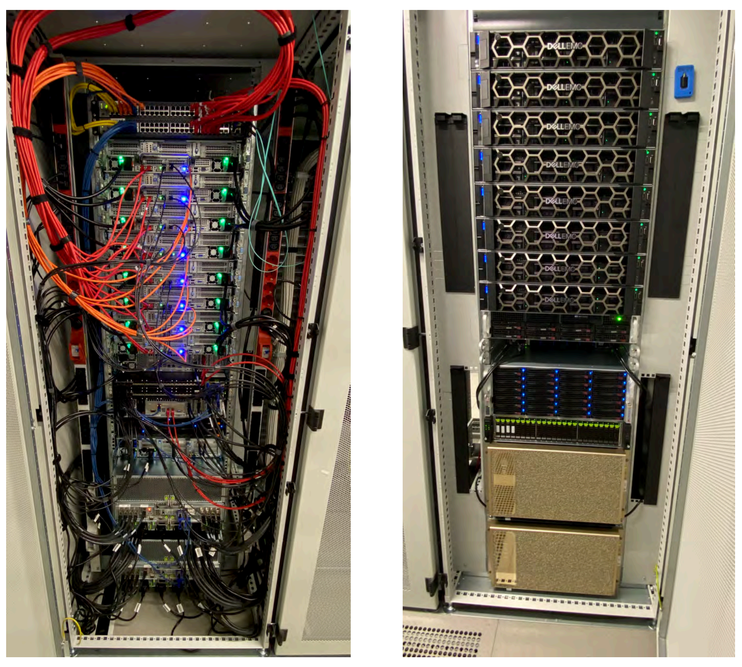

Abbildung 1: Zwei Nvidia DGX Server und acht Dell Server – in der Mitte Backup, Flash- und Login Server, sowie Network Switches

Insgesamt verfügt die Hochschule über 2 DGX A100 Server mit insgesamt 20 NVIDIA A100 GPUs und vier 64-Kern AMD CPUs.

CO2-FUSSABDRUCK

Die intensive Nutzung von Hochleistungsrechnern hat ihren Preis: Das Trainieren und die Verwendung großer KI-Modelle haben einen nicht zu vernachlässigenden Energiebedarf. So schätzt man z.B.

	CO2 Ausstoß
Google Anfrage	0,2 g CO_2
ChatGPT Query	4,3 g CO_2

Geschätzter CO_2-Ausstoß von Google und ChatGPT [4, 5, 3]

Google verzeichnet etwa 8,5 Milliarden Suchanfragen pro Tag, ChatGPT verzeichnet etwa 50 Millionen Besuche pro Tag.

Zum Vergleich: Unternehmens-IT verantwortet in Deutschland ca. 17 Megatonnen (Mt) CO_2-Emissionen.

Wachsende Emissionen im Konflikt zum 1.5°-Ziel

Laut der Internationale Energieagentur (IEA) müssen die Emissionen in dem Sektor Elektizität und Heizung, zu dem auch die Erzeugung von Strom für Hochleistungsrechner zählt, bis 2050 von 13.821 Mt CO_2 auf -369 Mt CO_2 reduziert werden [2].

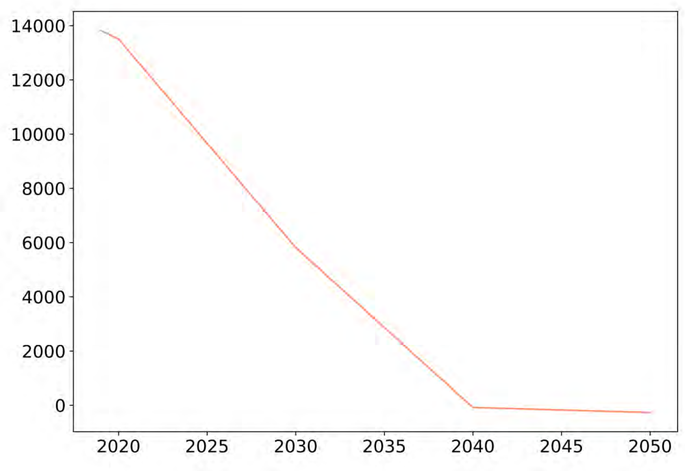

Abbildung 2: IEA Net Zero by 2050 Pfad für den Sektor Elektizität und Heizung

Dieses Ziel steht im Konflikt mit dem wachsenden Einsatz von Künstlicher Intelligenz - daher ist ein überlegter und bewusster Umgang erforderlich.

Durch die Messung und Transparenz der Emissionen bei der Cluster-Nutzung können Forschende an der Hochschule einen Beitrag zur Sensibilisierung für die Auswirkungen von KI auf das Klima leisten.

CO2-EMISSIONEN MESSEN

Eine genaue Schätzung von Emissionen ist erforderlich, um gezielt Steuerungsimpulse zur Reduzierung, etwa durch *Green Coding*, zu setzen.

Mit Tools wie dem *experiment-impact-tracker* [1], welches auf dem HPC Cluster zur Verfügung steht, kann man die CO_2 Emissionen von KI-Forschung einfach und hinreichend genau quantifizieren. Dies geschieht in zwei Schritten:

Messung der Energiekonsumptiom während der Laufzeit

Die Energiekonsumption in Rechenzentren ist wie folgt verteilt:

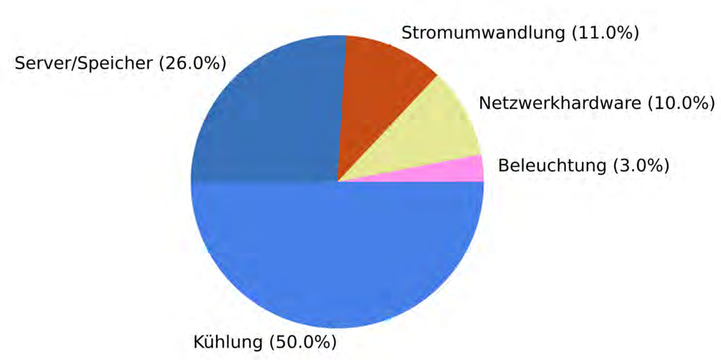

Abbildung 3: Quellen des Stromverbrauchs in Rechenzentren

Im ersten Schritt wird die Energiekonsumption aggregiert:

$$e_{total} = \text{PUE} \sum_{p} (p_{\text{dram}} e_{\text{dram}} + p_{\text{cpu}} e_{\text{cpu}} + p_{\text{gpu}} e_{\text{gpu}})$$

p_{resource} sind die prozentualen Anteile der einzelnen Systemressourcen, e_{resource} ist die Energiekonsumption dieser Ressource und PUE ist die Leistungseinheitseffizienz.

Berechnung der Emissionen

Im zweiten Schritt werden auf Grundlage dieser Energiekonsumption dann die resultierenden Emissionen berechnet. Diese ergeben sich u.a. aus dem Standort des Rechenzentrums.

QUELLEN

[1] Peter Henderson et al. *Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning*. 2022. arXiv: 2002.05651.

[2] IEA. *Net Zero by 2050*. 2021. DOI: https://www.iea.org/reports/net-zero-by-2050.

[3] David Patterson et al. *The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink*. Feb. 2022. DOI: 10.36227/techrxiv.19139645.v2.

[4] Emma Strubell, Ananya Ganesh, and Andrew McCallum. *Energy and Policy Considerations for Deep Learning in NLP*. 2019. arXiv: 1906.02243.

[5] BigScience Workshop. *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. 2023. arXiv: 2211.05100.