ESCADE:
Energy-efficient Artificial Intelligence for Cost-effective and Sustainable Data Centers

Sabine Janzen[1] | *Hannah Stein*[1,2] | Katharina Trinley[1,2] | Cicy Agnes[1] | Vaibhav Jain[1] | Karan Rajshekar[1]| Nirav Shenoy[1] | Anika Rusch[1] | Sujatro Ghosh[1,] | Wolfgang Maass[1,2]

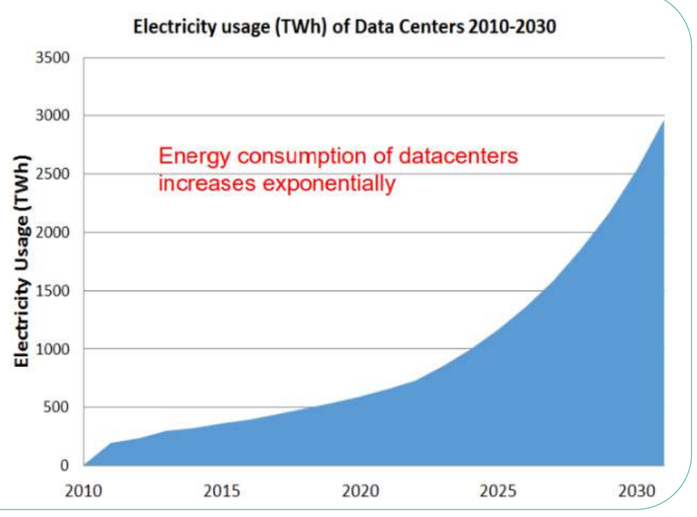
[1]*German Research Center for Artificial Intelligence, Germany*
[2]*Saarland University, Germany*

Optimizing AI workloads through compression, neuromorphic computing, and intelligent energy analytics

Motivation

Exponential increase of power of AI computing

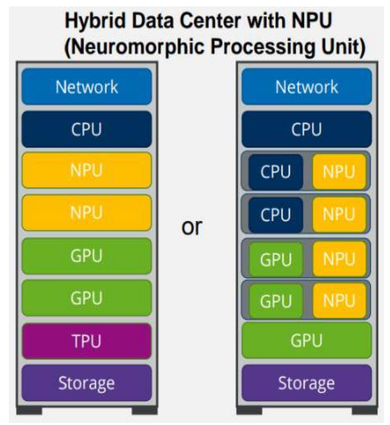
- By 2030: 13% of global energy consumption in data centers [1]
- No unified approach to Improve energy efficiency across various layers of ML stack



Approach & Method

ESCADE enables sustainable AI by combining:

- **AI model compression** (e.g., NAS, KD)
- **Neuromorphic hardware** for energy-efficient inference [5]
- A **decision-support tool** to balance **cost**, **performance**, and **carbon impact of AI operations**

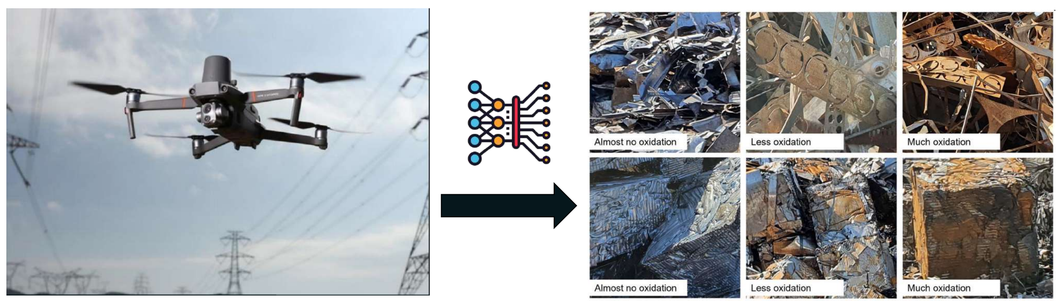


Objectives

- Significant **reduction of energy consumption** of data centers
- Improve the **cost-sustainability trade-off of AI applications** in data centers
- Combine **neuromorphic hardware**, **AI compression techniques** and AI-based energy management for sustainable data centers

Use Cases

Steel Industry



Large vision models used for real-time steel scrap sorting [2]

Consume >650,000 kWh/year
~€330,000 / deployment [3]

Through model compression

50% energy and inference savings,
~334,500 kWh and €167,250 / deployment

NLP for Software Tools

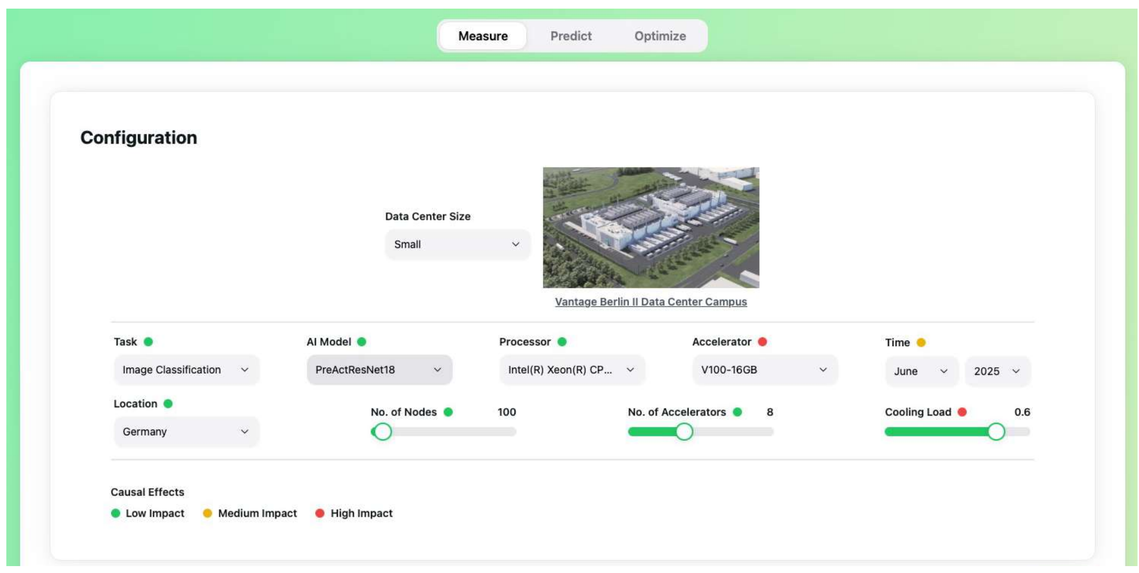
Fine-tuning NLP
~18 GPU hours
378 kWh/year/model

50% lower training energy
80% lower inference energy
189 kWh/year/model [4]

EAVE: Energy Analytics for Cost-effective and Sustainable Operations

EAVE supports energy-aware AI deployment through three integrated modules

- **Measure:** Cross stack sustainability data to track energy use, CO_2 emissions, and operational costs, including Power Usage Effectiveness (PUE) [6].
- **Predict** uses machine learning to forecast and recommend optimal spatio-temporal deployment configurations.
- **Optimize** multi-objective optimization to select optimal baseline vs. compressed AI models



Configuration panel in the EAVE system

Impact

- Embeds sustainability into the **entire AI model lifecycl**e, from design to deployment
- Bridges socio-technical silos by **combining AI engineering, decision support, and environmental responsibility**

Supports **UN Sustainable Development Goals**:

- Goal 9: Industry, Innovation, and Infrastructure
- Goal 12: Responsible Consumption and Production
- Goal 13: Climate Action

→ Reference architecture for future **energy-efficient, green data centers**

Supported by:

Federal Ministry for Economic Affairs and Energy

on the basis of a decision by the German Bundestag

Project Partners:

DFKI, TU Dresden, Univ. Bielefeld, SHS, NT.AG, SEITEC, Salzburg Research

Project Duration:

01.05.2023 – 30.04.2026

References:
1. Andrae, A. S. G., & Edler, T. (2015). On global electricity usage of communication technology: Trends to 2030. Challenges, 6(1), 117–157.
2. Schäfer, M., Faltings, U. & Glaser, B. DOES - A multimodal dataset for supervised and unsupervised analysis of steel scrap. Sci Data 10, 780 (2023).
3. A. Fu, M. S. Hosseini, K. N. Plataniotis, Reconsidering co2 emissions from computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2311–2317
4. Intel, Intel advances neuromorphic with loihi 2, new lava software framework and new partners, 2022. URL: https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-neuromorphic-loihi-2-lava-software.html#gs.fmm9op.
5. T. Schoormann, G. Strobel, F. Möller, D. Petrik, P. Zschech, Artificial intelligence for sustainability—a systematic review of information systems literature, Communications of the Association for Information Systems 52 (2023) 8.
6. N. Horner, I. Azevedo, Power usage effectiveness in data centers: overloaded and under-achieving, The Electricity Journal 29 (2016) 61–69.

Contact:

Dr.-Ing. Sabine Janzen
sabine.janzen@dfki.de
www.escade-project.de

EAVE Demo

YouTube

Deutsches Forschungszentrum
für Künstliche Intelligenz
German Research Center for Artificial Intelligence

ESCADE: Energy-efficient Artificial Intelligence for Cost-effective and Sustainable Data Centers

Sabine Janzen[1] (sabine.janzen@dfki.de) | Hannah Stein[1,2] (hannah.stein@dfki.de)
[1]German Research Center for Artificial Intelligence, Germany
[2]Saarland University, Germany

Optimizing AI workloads through compression, neuromorphic computing, and intelligent energy analytics.

Motivation

Data centers are the backbone of digital transformation, especially in the age of artificial intelligence. However, their energy footprint is rapidly becoming unsustainable. Global energy consumption from data centers is expected to exceed 500 TWh annually by 2027 [1]. Training large AI models can emit as much CO_2 as the lifetime of multiple SUVs. With increasing regulatory pressure and energy costs, companies urgently need solutions that support both performance and sustainability.

The ESCADE project addresses this growing challenge by enabling sustainable, energy-aware deployment of AI models. It introduces a combination of AI compression methods, emerging neuromorphic hardware, and a decision-support system that helps organizations optimize the trade-off between cost, performance, and environmental impact.

Use Cases

Steel Industry

Steel production is inherently recyclable, but only 40% of steel today comes from scrap. One barrier is the lack of real-time, energy-efficient scrap sorting. Large-scale models like ResNet, used in computer vision for sorting, require over 650,000 kWh/year and cost over €330,000 in energy alone [2].

By compressing these models and using neuromorphic inference, ESCADE enables up to 50% energy savings and faster inference, improving classification quality and enabling higher reuse of scrap steel [3]. If widely adopted, this could translate to national energy savings of nearly **30 billion kWh** and €15 billion in economic benefit in Germany alone.

NLP for Software Tools

Customer support systems using natural language processing (NLP) often require fine-tuning or training of multiple topic extraction models. In practice, training, a single model consumes ~378 kWh/year, causing significant overhead when scaled across clients.

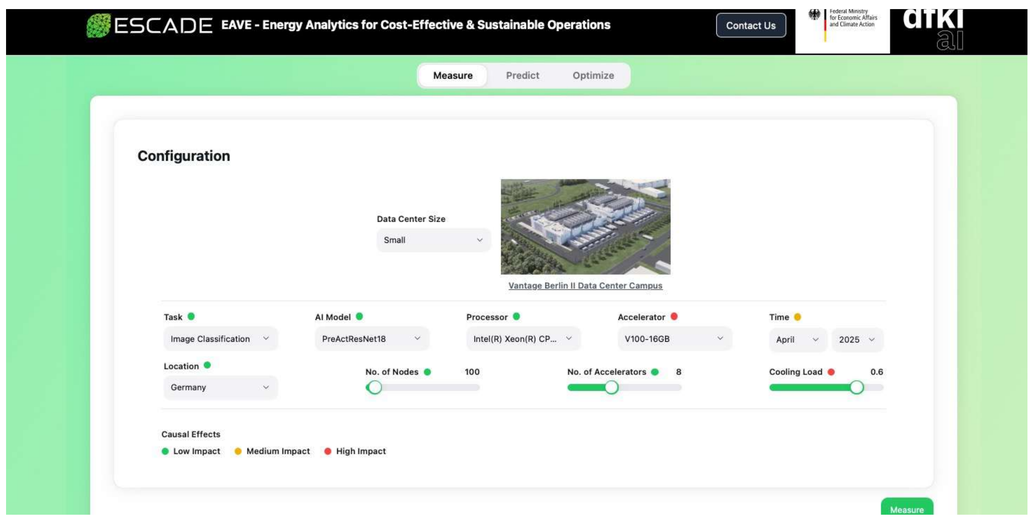
ESCADE reduces training energy by 50% and inference by 80% without performance degradation [4]. This enables scalable deployment of NLP models in business environmentswith potential savings of €95,000 per year for systems used by 1,000 customers.

ESCADE

EAVE Platform

A core outcome of ESCADE is the EAVE platform—Energy Analytics for Cost-effective and Sustainable Operations. EAVE supports energy-aware AI deployment through three integrated modules:

- **Measure** tracks energy use, CO_2 emissions, and operational costs, including Power Usage Effectiveness (PUE) and key drivers like hardware load and cooling [5].
- **Predict** uses machine learning to forecast energy and emissions across regions and seasons, recommending optimal deployment configurations.
- **Optimize** compares baseline and compressed AI models to highlight trade-offs in energy, cost, accuracy, and performance.



Configuration panel and the Optimize Component in the EAVE system (www.eave.dfki.de)

Impact

ESCADE advances the engineering of information systems by embedding sustainability into the AI model lifecycle. It aligns with CAiSE's goals of bridging silos in socio-technical systems by bringing together AI engineering, decision support, and environmental responsibility. The project supports multiple UN Sustainable Development Goals (9, 12, and 13) and serves as a reference architecture for next-generation, green data centers.

References:

1. Goldman Sachs, 2024. URL: https://www.goldmansachs.com/insights/articles/ai-to-drive-165-increase-in-data-center-power-demand-by-2030.
2. A. Fu, M. S. Hosseini, K. N. Plataniotis, Reconsidering co2 emissions from computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2311–2317.
3. Intel, Intel advances neuromorphic with loihi 2, new lava software framework and new partners, 2022. URL: https://www.intel.com/content/www/us/en/newsroom/news/intel-unveils-neuromorphic-loihi-2-lava-software.html#gs.fmm9op.
4. T. Schoormann, G. Strobel, F. Möller, D. Petrik, P. Zschech, Artificial intelligence for sustainability—a systematic review of information systems literature, Communications of the Association for Information Systems 52 (2023) 8.
5. N. Horner, I. Azevedo, Power usage effectiveness in data centers: overloaded and under-achieving, The Electricity Journal 29 (2016) 61–69.

Project Partners:
DFKI, TU Dresden, Univ. Bielefeld,
SHS, NT.AG, SEITEC, Salzburg
Research

Funded by:
BMWK – GreenTech Innovation Competition
Duration: 2023–2026
www.escade-project.de | www.eave.dfki.de