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] and Wolfgang Maass[1,2]

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

Abstract

Data centers play a central role in digital transformation, especially in the field of artificial intelligence (AI). However, their energy consumption is enormous, e.g., 20 billion kWh in Germany in 2024. At the same time, energy costs are rising and climate neutrality requirements are increasing. These factors pose major challenges for the sustainable and cost-effective operation of data centers. This paper introduces the ESCADE project (05/2023 - 04/2026), an ongoing research initiative funded by the German Federal Ministry of Economics and Climate Action, aiming to optimize the energy-efficiency of AI in data centers. AI compression techniques such as knowledge distillation, quantization and neural architecture search result in smaller, more energy-efficient AI models that deliver comparable performance. When combined with neuromorphic hardware, these models can achieve energy savings of up to 80%. The ESCADE consortium, a multidisciplinary collaboration of seven industry and academic partners, explores energy-efficient AI in two use cases: visual computing for scrap sorting in steel industry and natural language processing for software development. This paper provides a comprehensive overview of the ESCADE project, outlining its objectives, work packages, and anticipated outcomes. A central contribution is the introduction of first results in terms of the information system EAVE: Energy Analytics for Cost-effective and Sustainable Operations. By using AI-based analyses, EAVE optimizes the relationship between AI performance and operating costs of AI applications in data centers. The system measures and predicts the energy consumption, CO_2 emissions and operating costs of different AI model configurations, including hardware options. At the same time, it analyzes which factors significantly influence these values. This enables decision-makers to manage the operation of data centers in a data-based and efficient manner while meeting environmental targets.

Keywords

Data center optimization, Energy efficiency, Artificial Intelligence, Sustainability in IS, Decision Support Systems, Socio-technical systems

1. Introduction

By 2027, the energy consumption of data centers worldwide will reach 500 terawatt hours (TWh) per year – an increase of 38% since 2023 [1]. The growth in data centers and their capacities is driven by digitalization, Industry 4.0, cloud computing and artificial intelligence (AI); Nearly 70% of German companies use data centers for operation and development of business-critical IT applications (turnover of €12.5 billion in 2021) [2, 3]. However, storing, processing, transporting and using data consumes energy and releases considerable amounts of CO_2 [4, 5]. Training larger AI models can emit the equivalent of five SUV lifetimes (284 tons of CO_2). For companies that develop AI-based services for innovative business models, there are currently no methods that help them to manage the implementation of AI models including training and inference with respect to cost-effectiveness and sustainability. This "black box" problem is a key driver behind the fact that, according to the annual electricity report from the International Energy Agency (IEA), data center energy consumption could more than double by 2026, exceeding 1,000 TWh in a worst-case scenario [4]. By 2030, data centers and their hosted AI applications will account for 13% of global energy consumption [6]. Together with

RPEatCAiSE25: Research Projects Exhibition at the International Conference on Advanced Information Systems Engineering, June 16–20, 2025, Vienna, Austria

*Corresponding author.
sabine.janzen@dfki.de (S. Janzen); hannah.stein@dfki.de (H. Stein); katharina.trinley@dfki.de (K. Trinley); cicy.agnes@dfki.de (C. Agnes); vaibhav.jain@dfki.de (V. Jain); karan.rajshekar@dfki.de (K. Rajshekar); nirav.shenoy@dfki.de (N. Shenoy); anika.rusch@dfki.de (A. Rusch); sujatro.ghosh@dfki.de (S. Ghosh); wolfgang.maass@dfki.de (W. Maass)

rising costs for electricity, gas, mineral oil and coal, the ability of companies to operate data centers economically is threatened [2]; forcing data centers to reduce energy consumption while maintaining the same computing power. This trade-off and the demand for sustainable operation is currently being addressed by sourcing the energy used from renewable energy sources [7] and using modern, direct or indirect free cooling systems [8]. However, this does not change the ineffective energy balance of classic data centers and AI algorithms.

ESCADE is an ongoing research project funded by the German Federal Ministry of Economics and Climate Action (05/2023 - 04/2026) that aims to improve the cost-sustainability trade-off of AI applications in data centers[1]. By means of compression techniques for energy-efficient AI, neuromorphic chip technologies and energy analytics for decision makers, companies, data center operators and government institutions can move from a reactive to a proactive stance by forecasting and balancing AI performance, sustainability and economics. AI compression techniques such as knowledge distillation, quantization and neural architecture search result in smaller, more energy-efficient AI models that deliver comparable performance. When combined with neuromorphic hardware, these models can achieve energy savings of up to 80%. The project consortium consists of seven partners from industry and academia that take different roles in the project: data center provider (NT.AG), industry end user (SHS), system provider (Seitec), and research & development (DFKI, University Dresden, University Bielefeld, Salzburg Research). ESCADE investigates the application of energy-efficient AI and neuromorphic computing in two domains - scrap sorting in steel industry (visual computing) and natural language processing for software development - with the objective to demonstrate considerable energy-saving potential.

In this paper, we present an overview of ESCADE, including its objectives, work packages and expected outcomes. We discuss the role of information systems in ESCADE, focusing on the current state of work and first results in form of energy analytics for cost-effective and sustainable operations (EAVE). By using AI-based analyses, EAVE optimizes the relationship between AI performance and operating costs of AI applications in data centers. The system measures and predicts the energy consumption, CO_2 emissions and operating costs of different AI model configurations, including hardware options. At the same time, it analyzes factors that significantly influence these values. This enables decision-makers to manage the operation of data centers in a data-based and efficient manner while meeting environmental targets.

2. Project objectives

Objective of the research project ESCADE is the significant reduction of energy consumption of data centers by using cutting-edge hardware and software technologies to improve the ecological footprint of AI applications. Concerning the hardware focus is on the use of neuromorphic chip technologies, as these promise efficiency gains of up to 50% in training and up to 80% in inference of AI models. Concepts for integrating neuromorphic processing unit into classic GPU-based data centers, AI compression techniques and AI-based energy management services will help planning new data centers more sustainable, while existing data centers will be supported for hybrid operation. A sustainable, resource-efficient design of data centers and AI systems contributes directly to achieving the UN Sustainable Development Goals 9, 12 and 13. The concepts for cost-effective and sustainable data centers will be demonstrated in two use cases for measuring sustainability and economic efficiency.

Use case 1: Sustainable steel industry through energy-efficient AI
Problem statement: The energy consumption of large visual computing (VC) models such as ResNet can exceed 650,000 kWh annually [9], costing approximately €330,000 to train a single model. Such models are necessary for scrap sorting in steel industry. In contrast to plastic, steel is an ecologically sustainable material per se, enabling a completely closed-loop economy in the steel industry [10]. Producing one ton of steel from scrap consumes around 3.5 times less energy than conventional

[1] www.escade-project.de

production [11] (potential savings of 89.3 billion kWh per year for the German steel industry) [12]. However, steel scrap has only been used for 40% of steel products [13], as correct and efficient visual pre-sorting in real time is impossible or very resource-intensive [14].
Objective: ESCADE enables data centers to process large VC models up to 50% [15] more efficiently in terms of energy consumption and inference time. For a large VC model, this corresponds to savings of energy (- 334,500 kWh) and costs (- €167,250). It enables effective and efficient classification of scrap types during pre-sorting, with a side effect of increasing use of steel scrap (at least 20%, i.e. savings of nearly 30 billion kWh and approximately €15 billion in Germany).

Use case 2: Large, energy-efficient NLP models
Problem statement: For NLP-based software such as ticket systems, no "one-fits-all" AI models can be offered, e.g., for topic extraction. The training or fine-tuning of sometimes multiple models requires approximately 18 hours on an NVIDIA V100 GPU and 2 Intel Xeon 642 CPUs (378 kWh and 540 hours per year and model). Thus, training is the bottleneck of the entire business model with respect to latency (e.g., to react quickly to customer requirements), scalability and energy consumption.
Objective: Data centers implementing ESCADE concepts can achieve a reduction in energy consumption of 50% for training and 80% for inference of NLP models [16], i.e., energy consume of 189 kWh/year/model for the aforementioned setting [17]. Based on 1000 customers of a ticket system, this means a potential for cost savings of around €95,000 per year for training the models.

Table 1
Descriptive information about the research project ESCADE

Name	ESCADE: Energy-Efficient Large-Scale Artificial Intelligence for Sustainable Data Centers
Duration	01.05.2023 – 30.04.2026 (36 months)
Funded partners	German Research Center for Artificial Intelligence GmbH (DFKI) (Coordinator), Technical University of Dresden (TUD), University Bielefeld (UB), Stahl-Holding-Saar GmbH & Co. KGaA (SHS), NT Neue Technologie AG (NT.AG), SEITEC GmbH (SEITEC), Salzburg Research Forschungsgesellschaft m.b.H. (Salz)
Funding agency	German Federal Ministry for Economic Affairs and Climate Action (BMWK)
Funding program	GreenTech Innovation Competition - Digital Technologies
Project volume	5 Mio. €
Website	escade-project.de and eave.dfki.de

The project is structured into ten work packages (WP) (cf. Fig.1) that were distributed among the partners based on their areas of expertise (cf. Tab. 1). WP2 is led by DFKI and develops a framework to measure the sustainability of data centers and AI algorithms. On this basis, the framework is implemented in WP3 through open-source software modules that enable energy analytics (SEITEC). WP4, led by TU Dresden, energy-efficient AI algorithms for Use Case 1 and 2 are implemented on conventional an neuromorphic hardware. Also led by TU Dresden, WP5 pursues the design & implementation of sustainable AI data centers through the creation of reference architectures with neuromorphic hardware for "the world's most sustainable AI data center." WP6 (University Bielefeld) and WP7 (Stahl-Holding-Saar) focus on the implementation and evaluation of Use Case 1 (WP6) and 2 (WP7). WP8, Evaluation & Benchmarking (NT.AG), assesses energy efficiency gains of the new data center concepts and performs utility evaluations. WP9, Founding & utilization initiatives (eco2050), applies economic analyses and examines utilization concepts (e.g., startup creation) based on the project results. Within WP10, knowledge transfer and communication of the results through publications, standardization, and community building are conducted. Within WP1, the project is coordinated through DFKI.

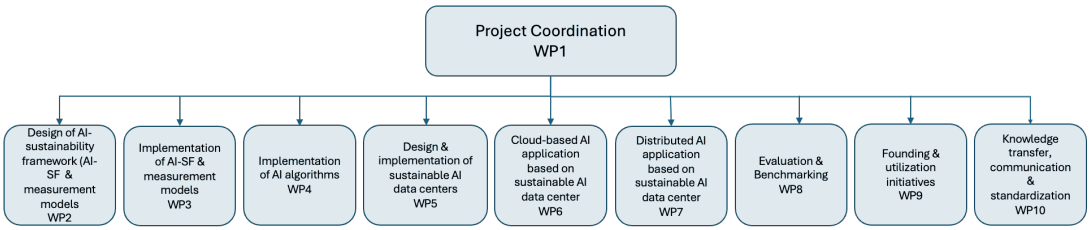

Figure 1: List of the workpackages in ESCADE.

3. Current status and intermediate results

The decision support system **EAVE - Energy Analytics for Cost-effective and Sustainable Operations** was developed - for data center providers as well as companies training and deploying AI models. EAVE[2] optimizes energy efficiency and cost-effectiveness of AI applications in data centers. It measures and predicts the energy consumption, CO_2 emissions, and operating costs of different AI model configurations, including hardware options (cf. Fig. 2). At the same time, it analyzes which factors significantly influence these values. EAVE consists of three main modules: **Measure, Predict** and **Optimize**.

The **Measure module** calculates operational energy costs, energy consumption, CO_2 emissions, as well as the Power Usage Effecitveness (PUE) in terms of an energy analytics summary based on hardware configurations, AI model, time and location. The PUE value represents an indicator for energy efficiency by providing the ratio of total facility energy to IT equipment energy [18]. The operational costs are calculated by using historical energy price data of the chosen location. In addition, EAVE provides a causal factor analysis [19] to quantify which variables (e.g., accelerator usage, cooling load) affect the PUE.

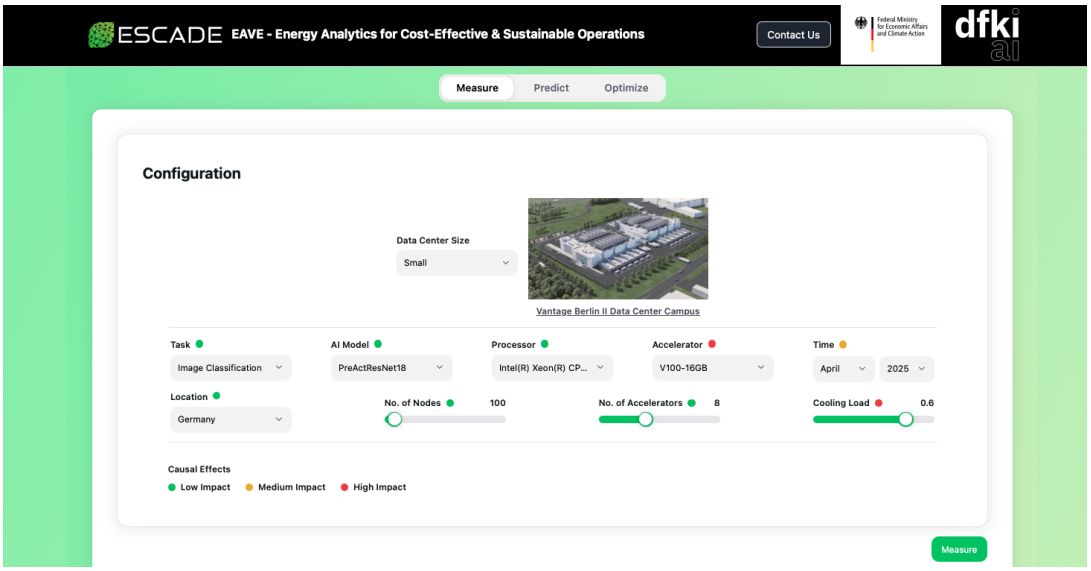

Figure 2: Configuration panel in the EAVE system

Based on the Measure module, the **Predict module** performs spatiotemporal optimization for the given data center configuration. It predicts the most efficient combination of location and time of year to minimize energy consumption, operational costs, and environmental impact, based on the hardware configuration selected in the Measure module. The module currently supports predictions across Germany, France, Netherlands, Italy, Poland, Austria, and the United States. Random Forest-based machine learning models [20] were trained to estimate key metrics, including energy costs, CO_2

[2] eave.dfki.de

emissions, and PUE value. The module displays a comparison between the initial setting (in the Measure module) and the predicted optimal setting with respect to costs, energy usage, PUE, and CO_2 emissions.

The **Optimize module** (Fig. 3) analyzes and optimizes AI model efficiency. Based on the initial hardware configuration, different baseline AI models (e.g., Vision Transformer (ViT)) can be selected and analyzed (costs, energy consumption, accuracy, model size and CO_2 emissions). The AI compression techniques (cf. Fig. 3) lead to more energy- and cost-efficient models. The techniques include Knowledge Distillation [21, 22], which transfers knowledge from a large model to a smaller one; Quantization [23], which reduces the precision of model parameters to decrease memory and energy usage; Neural Architecture Search [24], that automates the design of more efficient AI architectures. The Optimize module compares the key metrics of the AI baseline models with the compressed AI models. It provides a visualization of Pareto-optimal models depending on model size, costs, accuracy, and inference time (cf. Fig. 3).

EAVE was implemented using Python, with React (frontend) and FastAPI (backend).

Figure 3: Optimize Component in the EAVE System

4. Relevance of project for CAiSE

The research project ESCADE directly contributes to the CAiSE community by addressing a critical challenge in modern information systems (IS) engineering: the sustainable and cost-effective operation of AI-driven applications. As digital transformation accelerates, IS increasingly rely on energy-intensive AI models and infrastructure. ESCADE advances the field through the development of energy analytics for IS, a novel capability that enables decision-makers to balance performance, sustainability, and economics both before and during system operation. This aligns with CAiSE topics of interest, including data-driven decision support, system optimization, and the socio-technical impacts of information systems. By embedding sustainability metrics such as CO_2 emissions and energy consumption into AI model lifecycle management, ESCADE expands the conceptual and technical foundation for engineering responsible, energy-aware information systems. Furthermore, the application of AI compression techniques and neuromorphic hardware introduces new design paradigms for creating energy-efficient information systems with applications beyond the two domains initially studied. This aligns with recent directions in the IS community to systematically embed sustainability goals within AI-based systems [17].

A central element of this contribution is the EAVE platform, which integrates decision support capabilities into the system lifecycle and enables organizations to account for cost, performance, and sustainability factors during both the design and deployment of AI-driven information systems. EAVE exemplifies the project's effort to operationalize sustainability within IS engineering processes, for example through AI compression techniques. By combining expertise from data center operations,

AI optimization, and decision support, ESCADE provides a multi-layered IS architecture that directly supports the CAiSE 2025 theme, "Bridging Silos." Its emphasis on interpretable analytics, cross-domain applicability, and socio-technical trade-offs reflects the CAiSE community's broader vision of designing systems that are not only intelligent but also responsible and resilient.

5. Conclusion and future work

The growing energy demands of data centers, driven by digital transformation and the advancement of AI, poses serious challenges to sustainability and economic efficiency. ESCADE addresses this critical issue by developing methods and tools that optimize the energy consumption of AI applications without performance losses. Through the implementation of AI compression techniques, neuromorphic hardware, and energy analytics, ESCADE enables stakeholders to shift from reactive to proactive strategies in managing AI workloads. The initial results of the decision support system EAVE illustrate the potential of data-driven energy analytics to reduce operational costs and CO_2 emissions in data centers. By quantifying the trade-offs between model performance, hardware choices, and environmental impact, EAVE empowers decision-makers with actionable insights for sustainable AI deployment.

Future work will extend the current system with real-time monitoring capabilities, training and inference times in data centers, and integrate additional AI use cases beyond the initial domains of VC and NLP. Further research is needed to generalize the findings for hybrid and pure neuromorphic hardware settings and related AI workloads. As energy costs and sustainability targets continue to rise in importance, ESCADE aims to become a blueprint for energy-conscious AI infrastructure in private and public sectors.

Acknowledgments

This work was partially funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) under the contract 01MN23004A.

Declaration on Generative AI

During the preparation of this work, the author(s) used Generative AI tools such as GPT-4 and Grammarly to assist with grammar correction, spelling, and occasional paraphrasing during the writing process (W). No generative AI tools were used to generate figures, tables, or scientific results. The research also involved experimentation with pretrained AI models (e.g., Vision Transformers and DeiT), and code assistance using tools such as ChatGPT or similar models may have been used for non-critical scripting tasks (C+E). All AI-assisted outputs were critically reviewed, validated, and edited by the author(s), who take full responsibility for the content of this publication.

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] ScaleUp Technologies, Bitkom study 2022: Data centers in germany, 2022. URL: https://www.scaleuptech.com/en/blog/bitkom-study-computing-centers-in-germany-2022/.

[3] Office of Technology Assessment at the German Bundestag (TAB), Energy consumption of ict infrastructure in germany, 2021. URL: https://www.tab-beim-bundestag.de/english/projects_energy-consumption-of-ict-infrastructure.php.

[4] International Energy Agency, Electricity 2024: Analysis and forecast to 2026, 2024. URL: https://iea.blob.core.windows.net/assets/6b2fd954-2017-408e-bf08-952fdd62118a/Electricity2024-Analysisandforecastto2026.pdf.

[5] United Nations Environment Programme, How artificial intelligence is helping tackle environmental challenges, 2022. URL: https://www.unep.org/news-and-stories/story/how-artificial-intelligence-helping-tackle-environmental-challenges.
[6] A. Liebl, M. Ballweg, M. Wehinger, T. Rückel, How to leverage ai to support the european green deal, 2022. URL: https://aai.frb.io/assets/logos/AppliedAI_SYSTEMIQ_Whitepaper_ClimateAI.pdf.
[7] International Energy Agency, Data centres and data transmission networks, 2022. URL: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks.
[8] U.S. Department of Energy, Best practices guide for energy-efficient data center design, 2011. URL: https://www.energy.gov/femp/articles/best-practices-guide-energy-efficient-data-center-design.
[9] 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.
[10] M. Sahoo, S. Sarkar, A. C. Das, G. G. Roy, P. K. Sen, Role of scrap recycling for co2 emission reduction in steel plant: a model based approach, Steel research international 90 (2019) 1900034.
[11] EuRIC aisbl, Metal recycling factsheet by euric, 2020. URL: https://circulareconomy.europa.eu/platform/en/knowledge/metal-recycling-factsheet-euric.
[12] C. Broadbent, Steel's recyclability: demonstrating the benefits of recycling steel to achieve a circular economy, The International Journal of Life Cycle Assessment 21 (2016) 1658–1665.
[13] C. Friedl, Schrott vor steiler karriere, 2021. URL: https://www.recyclingnews.de/rohstoffe/schrott-vor-steiler-karriere/.
[14] R. J. Compañero, A. Feldmann, A. Tilliander, Circular steel: how information and actor incentives impact the recyclability of scrap, Journal of Sustainable Metallurgy 7 (2021) 1654–1670.
[15] 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.
[16] D. Schmidt, G. Koppe, M. Beutelspacher, D. Durstewitz, Inferring dynamical systems with long-range dependencies through line attractor regularization, arXiv preprint arXiv:1910.03471 (2020).
[17] 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.
[18] N. Horner, I. Azevedo, Power usage effectiveness in data centers: overloaded and underachieving, The Electricity Journal 29 (2016) 61–69.
[19] A. Sharma, E. Kiciman, Dowhy: An end-to-end library for causal inference, arXiv preprint arXiv:2011.04216 (2020).
[20] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[21] J. Gou, B. Yu, S. J. Maybank, D. Tao, Knowledge distillation: A survey, International Journal of Computer Vision 129 (2021) 1789–1819.
[22] M. Phuong, C. Lampert, Towards understanding knowledge distillation, in: International conference on machine learning, PMLR, 2019, pp. 5142–5151.
[23] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, K. Keutzer, A survey of quantization methods for efficient neural network inference, in: Low-power computer vision, Chapman and Hall/CRC, 2022, pp. 291–326.
[24] B. Zoph, Q. V. Le, Neural architecture search with reinforcement learning, arXiv preprint arXiv:1611.01578 (2016).