ESCADE

Energy-Efficient Large-Scale Artificial Intelligence for Sustainable Data Centers

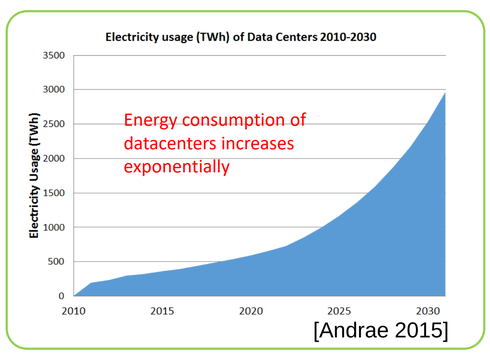
Bernhard Vogginger[1], David Kappel[2], Ulrike Faltings[3], Michael Schäfer[3,4], Andreas Hantsch[5,6], Sebastian Gawron[7], Dusan Dokic[8]

1 TU Dresden **2** Institute for Neural Computation, RU Bochum **3** SHS - Stahl-Holding-Saar GmbH & Co. KGaA, **4** KTH Royal Institute of Technology, Department of Materials Science and Engineering **5** eco2050 Institute for Sustainability GmbH **5** Hantsch Sustainability Consulting, **7** NT Neue Technologie AG, **8** German Research Center for Artificial Intelligence (DFKI)

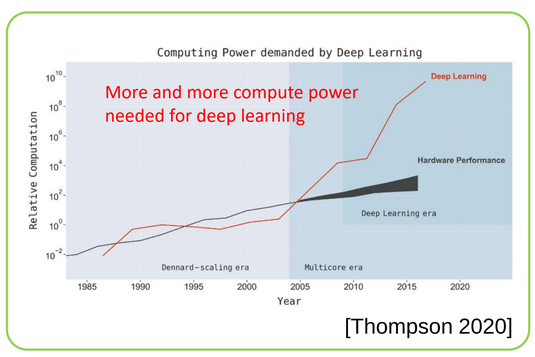
AI Datacenters and Sustainability

Exponential increase of power for AI computing

- **Energy use** in German datacenters **increased by more than 2x** from 5.8 bn in 2010 to 16 bn kWh in 2020.
- Forecast 2030: **13% of global energy consumption in datacenters**, big part for AI
- **Compute power needed** for digital technologies, especially **AI**
- **Large deep learning models** need **more and more compute power** for training and inference



[Andrae 2015]



[Thompson 2020]

Challenges:

- Rising demand for compute power jeopardizes sustainability goals
- Existing hardware (CPU/GPU) not efficient enough to reduce resource usage
- Also consider other resources: water, greenhouse gases, rare earth elements

ESCADE Overview

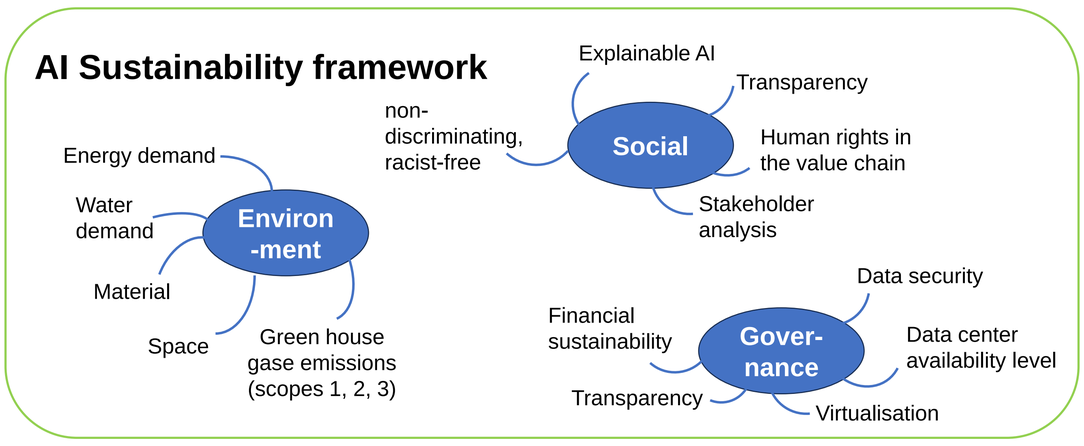
Project facts:

- **GreenTech Innovation Competition** by German Federal Ministry for Economic Affairs and Climate Action (**BMWK**)
- 5 million €
- May 2023 – April 2026

Goals

- Design **world‘s most sustainable AI data centers** on the basis of **neuromorphic hardware** (NPU, SpiNNaker2)
- **AI sustainability framework** to measure ressource usage of entire ML lifecycle (Development, training, inference) in datacenter
- Develop **End-to-End Sustainable AI solutions** for 2 use-cases

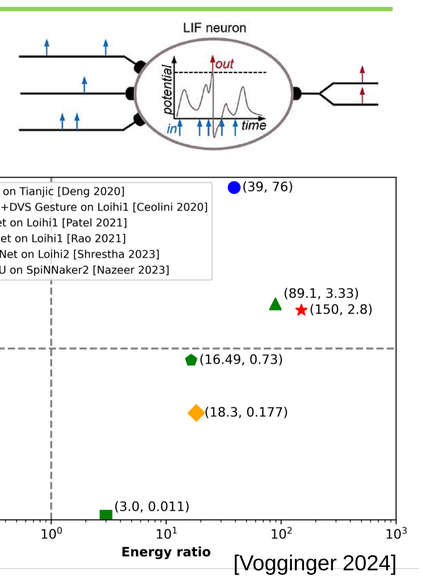
AI Sustainability framework



SpiNNaker2 Neuromorphic Hardware at TUD

Neuromorphic Chips (NPUs)

- **Neuro-inspired processors** for spiking neural networks
- Efficiency through **asynchronous, event-based processing**
- Intel Loihi and **SpiNNaker2**: **10-100x faster and more efficient than CPU/GPU** for DNN
- Neuromorphic approach applicable to many of deep learning models:
  - Image processing with CNN
  - NLP with RNN (sLSTM or EGRU)
  - Spatiotemporal pattern recognition
- *Is neuromorphic computing applicable to large-scale DL models?*
- *What is needed for the successful integration of NPUs into data centers?*

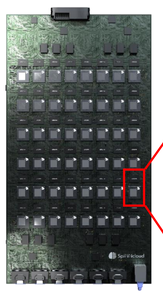


[Vogginger 2024]

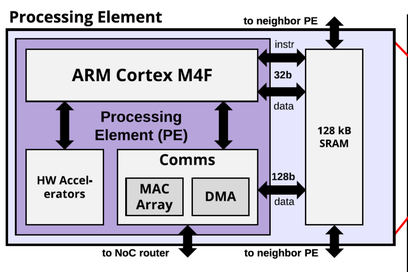
SpiNNcloud neuromorphic supercomputer

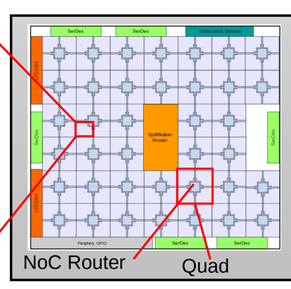
5 million core machine from 2024 in Dresden:
8 racks with 90 48-node boards each

48-node board



Node





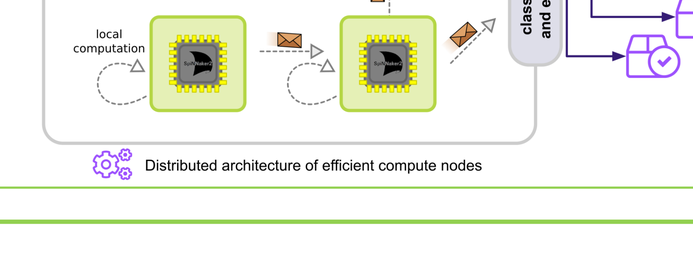
SpiNNaker2 Chip

- 152 PEs with Arm Cortex-M4F and DNN accelerators
- SpiNNaker router
- 6 chip-to-chip links
- GF 22FDX
- 103 mm²
- 0.5 V ABB IP from Raciycs

Industrial Use-Cases

1. Visual Computing for steel industry

- Steel scrap is a 100% circular material
- Scrap sorting can help optimize scrap usage
- Efficient classification of scrap is needed for green steel
- Requires realtime visual object identification
- Avoid rebound effects from energy-intensive DL applications
- Goal: **Reduce energy consumption by 50% for inference** with distributed hardware (Edge-to-cloud) and combining NPU, GPUs and CPUs



2. Efficient training of NLP models for digital industry

- Ex. automic ticket system: no „one-fits-all“ solution for all customers
- Regular training or finetuning required. Costs for training can become bottleneck of the business model
- Goal: **Reduce energy for training by 50% and for inference by 80%** by efficient, event-based NLP Models on neuromorphic hardware
- **Efficient algorithm:** apply principles of event-based communication and lazy computation in RNN [Subramoney 2023] to NLP models (Transformers)
- **Efficient hardware:** implement on SpiNNaker2 (recent work by [Nazeer 2023] requires 18x less energy for language modelling than NVIDIA A100)

Contact:

Email: pm-escade@dfki.de

https://escade-project.de LinkedIn

Partners:

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, NT.AG, SEITEC SEIFERTH TECHNOLOGIES, TECHNISCHE UNIVERSITÄT DRESDEN, RUB, salzburgresearch, SHS - STAHL-HOLDING-SAAR

Subcontractors:

ALEPH ALPHA, eco2050, SpiNNcloud, elevait

References:

- Andrae, Anders SG, and Tomas Edler. "On global electricity usage of communication technology: trends to 2030." *Challenges* 6.1 (2015): 117-157.
- Nazeer, Khaleelulla Khan, et al. "Language Modeling on a SpiNNaker 2 Neuromorphic Chip." *arXiv preprint arXiv:2312.09084* (2023).
- Subramoney, Anand, et al. "Efficient recurrent architectures through activity sparsity and sparse back-propagation through time." 11th ICLR (2023)
- Thompson, Neil C., et al. "The computational limits of deep learning." arXiv preprint arXiv:2007.05558 (2020).
- Vogginger, Bernhard et al. „Neuromorphic hardware for sustainable AI data centers.“ *arXiv preprint arXiv:2402.02521 (2024).*

Sponsored by:

Gefördert durch:

Bundesministerium für Wirtschaft und Klimaschutz

DLR Projektträger

aufgrund eines Beschlusses des Deutschen Bundestages