NOKIA BELL LABS

Advancing AI: Sustainability for networks

Reducing energy consumption of AI in networks – a pragmatic approach

White paper

The rapid, global adoption of AI and, especially, today's large language models (LLMs) is evidenced by the explosive growth in usage and funding it has created. However, AI adoption is also posing environmental and economic challenges due to soaring energy consumption for training and inference. This paper addresses the need for energy efficiency within the telecommunications and networking sectors, where AI is foundational to 6G and network autonomy. We introduce the Energy-efficient AI for Networks Guide (EA4NG), a pragmatic, three-step framework that ensures AI for networks minimizes its energy footprint and maximizes its energy handprint. Using systematic optimization techniques like pruning, quantization and specialized hardware as well as mandatory consumption monitoring, telecommunication providers and AI and data center operators can achieve sustainable AI for networks. The paper advocates for a multi-pronged strategy encompassing brain-inspired AI paradigms and hardware-software co-creation to achieve ambitious energy reduction goals, securing the future profitability and sustainability of AI deployments.

Anne Lee, Gurudutt Hosangadi, Joachim Wabnig, Marc-Olivier Buob, Mikko Honkala, and Sean Kennedy

Contents

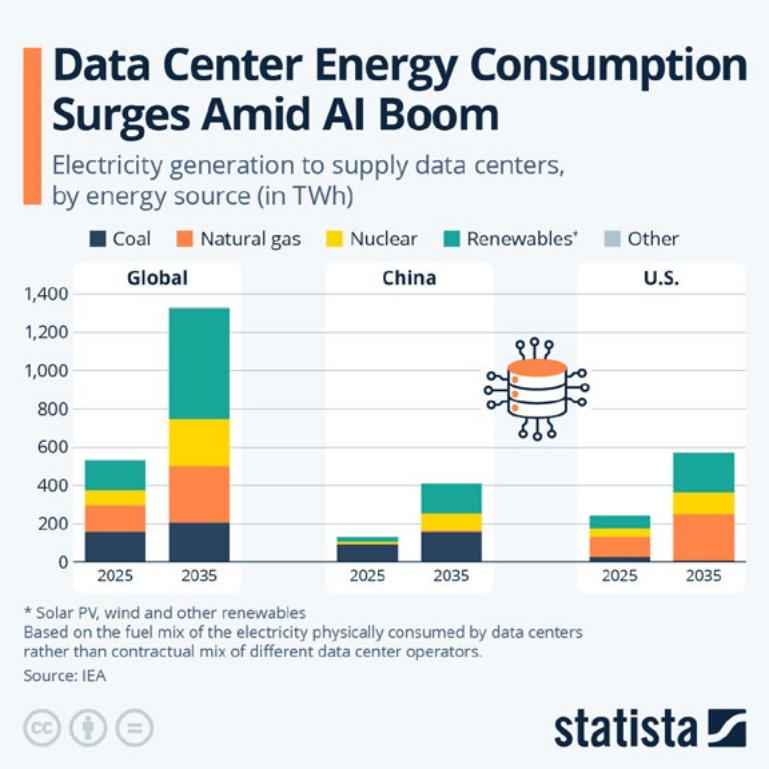
Introduction

The AI revolution is well underway. The recent advent of large language models (LLMs) and their use in chatbots, coding and professional functions have seen user adoption skyrocket, with ChatGPT reaching 800 million weekly users by the end of 2025 [1]. The appetite for using AI in a personal and professional context and the hope of ever-increasing AI capabilities has led to a subsequent explosive growth in venture funding and re-orienting of research initiatives across industries towards AI. To meet demand for the computing power needed to run AI models, data center capacity is rapidly expanding around the globe.

The power of AI demonstrated by these models comes with a severe cost to natural resources: use of rare earth metals in processors, use of fresh water for cooling the data centers, and, most of all, energy use—first for training the AI models and second for serving them to the end users. Energy use translates directly into operating expenses through electricity costs. This means that AI models' energy use directly impacts the capability, feasibility and profitability of AI deployments.

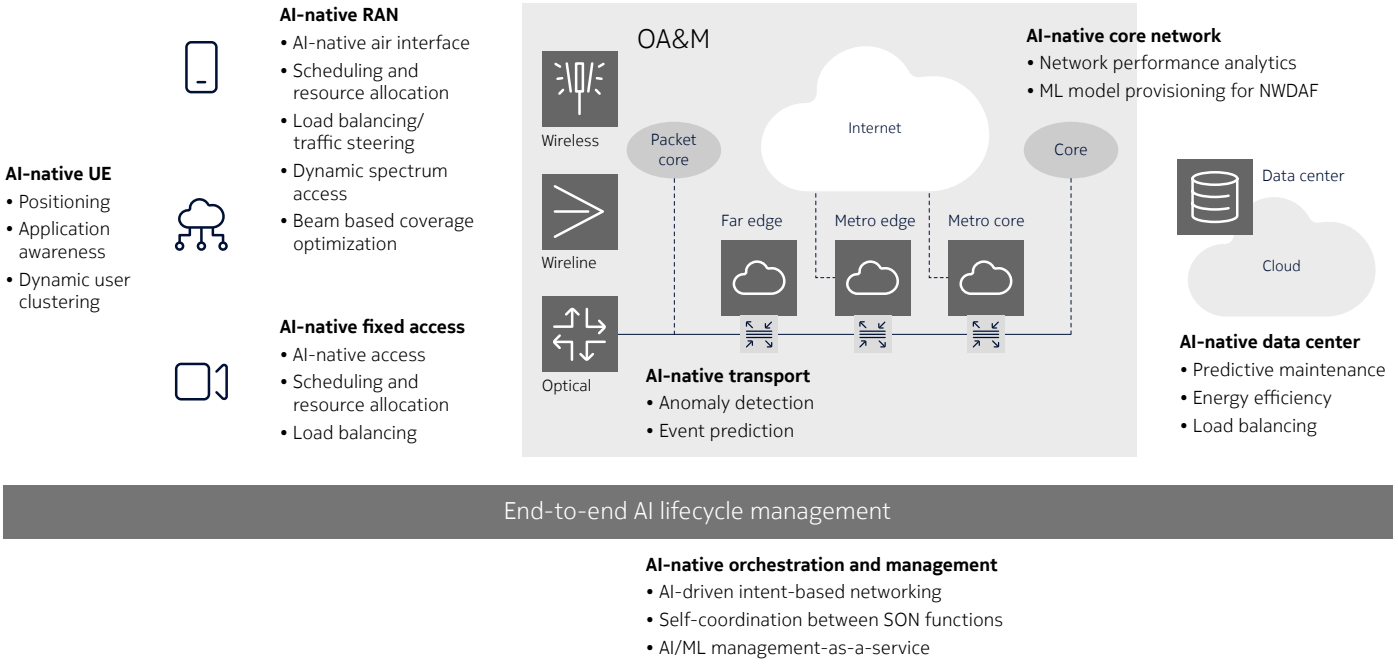
As an example, GPT–3.5, the model underlying the original ChatGPT, consumed ~1 GWh of energy during training, equivalent to the energy usage of an average American household over 120 years [2]. In addition, the amount of energy required for serving AI models is considerable: with 2.5 billion queries every day and an energy cost of at least 0.3 Wh per query, ChatGPT is estimated to consume about 275 GWh of electric energy per year [3, 4]. These numbers will only rise with the increase in capability and number of future AI models. The reaction to this projected growth in demand is an expansion of planned data center capacity worldwide; today, there are over 11,000 data centers [5]. The global energy consumption of data centers is predicted to grow from 500 TWh in 2025 to 1300+ TWh by 2035.

Figure 1. A summary of expected data center growth over the next decade, globally and in different regions along with the estimated energy mix used to power them [6]



In telecommunications, networking, and adjacent fields (industrial automation, virtual and augmented reality), there is a desire to harness the power of AI for automation and improve performance and value creation through new functionality. As shown in figure 2, AI is under consideration for use in many parts of the communications chain.

Figure 2. AI-native end-to-end autonomous network system



The major standardization body in mobile communications, 3GPP, reflects this by making both AI-nativeness and sustainability foundational to 6G [7]. We anticipate that wired networks will drive similar adaptation to achieve TM Forum's Level 5 autonomy [8]. The telecommunications and networking industry is keenly aware of energy costs, as well as latency, memory and power budget constraints in computing hardware. We must carefully examine these constraints and determine a balance between performance gains and energy costs before introducing AI models.

Energy is emerging as a major constraint for AI operations due to cost, energy supply and sustainability concerns. Further AI adoption and growth, particularly in the telecommunications sector, is only possible with a reduction in energy consumption.

Despite rapid advancements and increased accessibility of graphics processing units (GPUs), the need for energy-efficient AI in networks remains critical. Although GPU energy efficiency continues to improve, it is approaching its limits [9]. Moreover, Jevons paradox suggests that these gains will likely lead to higher energy consumption as AI scaling laws exploit the energy that is saved for further enhancements in accuracy and capabilities. Energy-efficient AI is key to driving superior performance and profitability. This paper highlights the feasibility and practicality of achieving optimal energy-efficient AI in networks through the strategies detailed in our proposed Energy–efficient AI for Networks Guide (EA4NG).

Goals

For the energy efficiency of AI for networks, our sustainability goals aim to substantially increase our positive impact (handprint) and reduce our energy consumption (footprint), while maintaining the same performance and accuracy of AI models during training and inferencing.

We want to build frugal AIs, therefore propose a pragmatic, systematic way for AI adoption to achieve best-in-class sustainable network solutions that will:

1. Ensure that AI is the right solution—we will present a guide for selecting the best path forward
2. Keep AI energy consumption to a minimum—for AI solutions, we will outline the engineering steps to reduce energy consumption, which will be illustrated by our own design choices in several use cases
3. Explore future technologies for AI energy efficiency—we will give an overview of the technologies and algorithms we believe will lead to significant reductions in AI energy use.

This paper will explain how to achieve these goals through our proposed Energy-efficient AI for Networks Guide (EA4NG).

Tenets

To achieve our goals, there are five tenets that will lead to the north star for AI energy efficiency in networks:

1. Smaller models—reducing the model size to the smallest needed for a particular task reduces energy requirements, and, for networks, this may also help reduce latency, which is important for some tasks
2. Learning efficiency—continual learning, fine-tuning, zero-shot, one-shot, or few-shot learning, transfer learning and knowledge distillation are methods to reuse or expand the knowledge of an AI model in ways that do not require full training or retraining
3. Sparse, event-driven computation—mimic brain activity where only a small fraction of neurons is active at any given time, leading to power consumption proportional to activity rather than total network size
4. Local computation—employ learning mechanisms that rely on local information, reducing the need for global communication and centralized control, which are energy intensive
5. Efficient compute and memory access—integrate memory and processing (in-memory computing) to reduce data movement, which is a major energy bottleneck in traditional computing architectures; also, leverage neuromorphic hardware designed for such operations.

State of the business

There are multi-dimensional approaches for reducing the energy footprint of AI in networks. While some approaches are used today, others are still being researched.

These approaches include both model selection and implementation. There are also various learning methods to choose from today. The energy requirements vary significantly among the model options listed below:

- Classical AI—decision trees, support vector machines (SVMs) and expert systems, etc

- Deep learning—deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs) and transformers, etc
- Brain-inspired AI—spiking neurons, liquid neural networks (LNNs) and baby dragon hatchling, etc.

For implementation, the techniques for reducing energy needs include:

- Model compression—smaller models use less energy for inferencing; techniques include pruning, for instance, with sparse neural networks (SNNs), quantization, knowledge distillation and low-rank factorization
- Energy-efficient hardware architectures are designed to run AI models for inferencing—technologies include in-memory compute and neuromorphic or analog compute (electrical, optical) chips
- Software architectural approaches are utilized both for training and inferencing-approaches include task specific model, modular design and mixture-of-experts (MoE).
- Training methods—these include various fine-tuning techniques and transfer learning.

Model compression

Minimizing the size of the model will result in reduced energy consumption. The following methods can be used to shrink the model [10]:

- Pruning is a technique that reduces the number of elements in AI models, which results in reducing redundancy and increasing sparsity
- Knowledge distillation is a technique that transfers knowledge from a large and complex teacher model to a small and simpler student model while maintaining the original level of performance
- Quantization is a technique that reduces the number of bits required to represent weights and activations; neural networks typically use 32-bit floating-point numbers for weights and activations, which can be quantized to 16-bit, 8-bit, 4-bit, and even 1-bit
- Low-rank factorization uses matrix and tensor decomposition to identify redundant parameters in an AI model, which results in decomposing large matrices into sets of smaller matrices—for dense layer matrices, storage and memory requirements are improved; for convolutional neural networks, factorization improves inference speeds.

Tools also exist to compress and accelerate AI models such as Neural Magic [11].

Hardware architectures

Various hardware approaches are being pursued to improve the performance of AI models. These are evolving technologies that are in various stages of research. They include (in the order of tech readiness with demonstrated gains):

1. Low-bit digital compute—already available in current products, e.g., as integer or low-precision floating-point compute and memory
2. In-memory computing—a specific hardware architectural approach that has the potential to reduce energy use because it eliminates the need to transfer data between the memory module and processor; several vendors are researching this area, including Nokia, IBM and Samsung [12, 13]
3. Analog electric chips—many companies are researching analog neural network chips; for example, IBM finds their analog approach to be 14x more energy efficient than the state-of-the-art approaches [14]
4. Neuromorphic chips—purposefully designed chips that run AI models, which are reaching the productization phase with companies such as IBM, Qualcomm and Samsung [15]

5. Analog photonic chips—a new chip known as the All-Analogue Chip Combining Electronics and Light (ACCEL) that uses photons for computing and transmitting information has reached a computing speed of 4.6 PFLOPS (peta-floating point operations per second) in lab tests; announced in 2023 by Tsinghua University, this chip was designed for vision tasks and is not yet on the market [16].

Software architectural approaches

There are many different software strategies to reduce the energy use of AI models. These include task-specific models and modular designs. Not all AI use cases, for instance, need general-purpose models. For example, AI for networks does not require the ability to write poems in the form of Shakesperean sonnets. Building models specific to a use case or task reduces the size of the model, hence, the energy required. Activating only the task-specific AI models when needed is analogous to activating only the neurons required in the brain, which makes the brain highly energy efficient.

There are several brain-inspired AI architectures that are open-sourced; they try to mimic the human brain in specific but varying ways:

1. Spiking neural network (SNN) architecture—instead of transmitting continuous activation values as in traditional neural networks, spiking neurons communicate by sending discrete "spikes" (brief electrical pulses) at specific times; example models include SpikeYolo [17], SpikingBrain [18], Thousand Brains Project [19] and Baby Dragon Hatchling [20]
2. Liquid neural nets (LNN) architecture—inspired by the "liquid" state of biological neural circuits, where a fixed, randomly connected recurrent network (the "liquid") transforms incoming signals into a high-dimensional, dynamic representation, which is then read out by a simpler, trainable layer; Massachusetts Institute of Technology (MIT) has spun off Liquid AI [21]
3. Hierarchical reasoning models (HRM)—inspired by the brain's hierarchical processing of information, HRMs break down complex problems into simpler, interconnected sub-problems, mimicking high-level cognitive processes like planning and decision-making; example models include Sapient Intelligence [22] and the tiny recursive model (TRM) [23].

TRM is an example of researchers pursuing ways to streamline existing and emerging architectural approaches (e.g., HRM) to reduce the size of the models while preserving high benchmark accuracy outcomes [23].

Efficient training methods

There are various ways to reduce energy consumption during training, for example:

- Reduce the amount of data required for training using zero-shot, one-shot, few-shot and transfer learning as well as fine-tuning to reduce the amount of energy required
- Enhanced backward propagation for DNNs, for instance, GreenTrainer, a fine-tuning technique, can save 64% of floating-point operations per second (FLOPS) without noticeable accuracy loss using adaptive backward propagation [24].

The co-design of training algorithm, model architecture and hardware

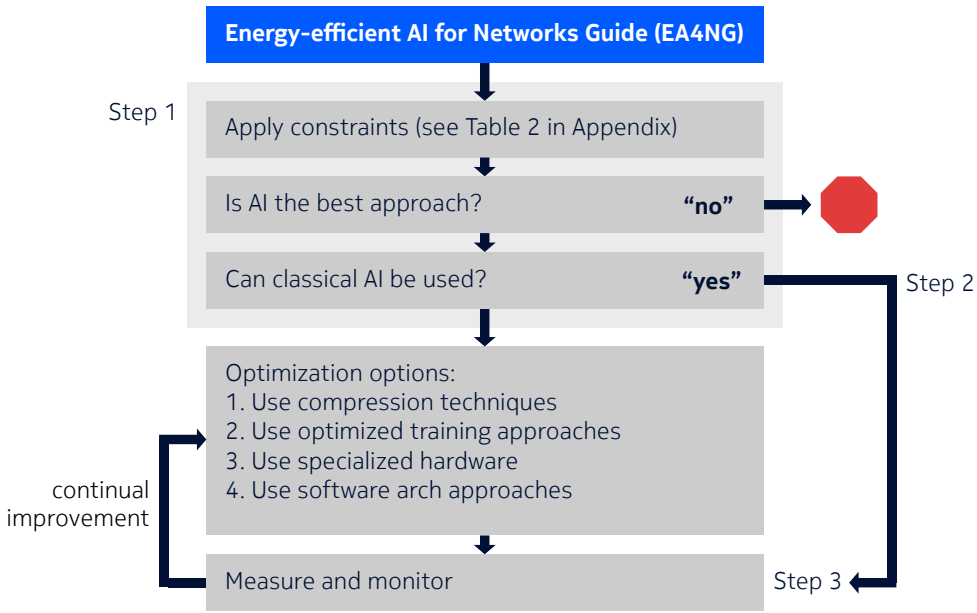
Current training methods and network architectures have been largely selected so that they train and run efficiently on GPUs. For other types of hardware, new methods will need to be developed to reap full benefits. One promising approach is to co-design the neural network architecture and the hardware jointly [25, 26].

Lessons learned:
Energy–efficient AI for networks guide (EA4NG)

The clear impact of the rapid adoption of AI and the resulting arms race is the soaring demand for energy. While solutions such as renewable energy and nuclear power plants can address some of the increase in demand, these are not sufficient since utilization of carbon–based energy is also rising. Reducing AI's energy consumption is imperative.

Many ideas and approaches are currently under research and implementation. These should make up a multi–pronged holistic solution since there is no single answer. Just as products are being designed now to be AI–native or AI–first, each AI–enabled product should also be designed at the beginning to be energy–efficient–first.

Figure 3. Energy–efficient AI for networks guide



An energy–efficient AI for networks guide has three essential steps followed by a set of best practice recommendations. See figure 3.

Step 1: Is AI needed, and, if so, then which one?

The first step is to determine whether AI is the right answer or approach for a network function. While AI is currently trendy, it is not always necessary or suitable for every situation. It is crucial for anyone considering AI to answer these questions:

a. Is there a real need [27, 28]?

b. Is it the best approach [28]? Does it meet the latency, memory, and power budget constraints?

c. Which type of AI is best? Neural, symbolic or neuro–symbolic AI? Language–based or non–language–based?

d. What is its footprint, and is it worth it? AI should be used only if its benefits offset its environmental impact [29, 30].

Ideally, one should consider every AI or non–AI–based approach to determine which solution has the best trade–off between energy efficiency and quality of the results [29, 31].

Related to asking "AI or not?" is the question of geography. What is the best location to deploy this AI implementation? The deployment aspects of AI (i.e., which country, which cloud, and which hardware architecture) must be seriously considered to minimize energy impact [30–32].

The entire toolchain must be considered, including the AI model, the underlying machine learning (ML) framework, the underlying graphics processing unit platform, and the underlying graphics processing unit. Candidate GPUs have an impact on energy efficiency. In AI, inferencing is both time– and space–constrained, but training can be relocated. Thus, training (and, ideally, inference) should be achieved with geographically-optimized energy–efficient hardware.

Step 2: AI energy optimization

Once the decision is made to use AI, we propose a systematic consideration and implementation of the following techniques to ensure energy minimization.

a. Use techniques like pruning, quantization and distillation to reduce the size of the models. Compression can affect the performance metrics of the models. Hence, there may be a trade–off. Understand and define what the acceptable performance thresholds are and compress accordingly.

b. Optimize the training process. Avoid full training or retraining when possible. Use transfer learning, few–shot learning and fine–tuning to minimize the amount of training required. Researchers are currently exploring efficient continual learning techniques to sustain the performance of trained models.

c. Use specialized hardware when possible. Usually, these will be more energy efficient than general–purpose processors. Neuromorphic chips and in–memory–compute show enormous promises.

d. Use brain–inspired models, if possible, that are sparse and mimic brain–like energy–efficient methods.

Step 3: Measuring and monitoring energy consumption

There is a need for universal AI benchmarks that assess energy efficiency. Without them, it is difficult to measure and compare progress. EcoLogits is a Python library that tracks the energy consumption and environmental footprint of generative AI models through application programming interfaces (APIs). It can be used to compare the environmental impact of various large language model–based tasks, based on the AI provider and the underlying model [33]. EcoLogits was created and is actively maintained by the GenAI Impact nonprofit [34].

Nokia proposes an environmental impact assessment of AI systems (ENVIAA) framework to transparently assess the environmental impacts of AI systems [29]. Nokia is also leading the EU Lighthouse Project on Sustainability in 6G, named SUSTAIN–6G [35]. This initiative will build upon the foundations of the Nokia–led flagship projects Hexa–X and Hexa–X–II. It will concentrate on the two key aspects, "Sustainable 6G" and "6G for sustainability," addressing all aspects of sustainability, including economic, environmental and societal.

Best practices recommendations

We recommend and propose using the following best practices for an energy–sustainable AI for networks solution:

- Design modular models to make it easier to share and reuse
- Design models for specific tasks rather than use general–purpose models
- Re–use trained models to save energy related to the learning process
- Use optimization techniques for training
- Optimize models to ensure energy–efficient products for inferencing
- Use hardware designed specifically for AI models
- Monitor energy consumption to assess the environmental impact.

And as much as possible, take geographic location into account when deploying AI to minimize energy consumption. See the Appendix for more details and network AI examples.

Conclusion

The rapid adoption of AI, particularly large language models, presents both immense opportunities and significant environmental and economic challenges due to soaring energy consumption. We have highlighted the critical need for energy efficiency within the telecommunications and networking sectors, where AI is foundational for future advancements like 6G and network autonomy. Nokia's pragmatic, three-step Energy-efficient AI for Networks Guide (EA4NG) provides a clear framework for minimizing AI's energy footprint and maximizing its positive impact. By systematically applying optimization techniques such as pruning, quantization, and leveraging specialized hardware, alongside mandatory consumption monitoring, telecommunications providers and AI/data center operators can achieve sustainable AI deployments. We advocate for a multi-pronged strategy, encompassing brain-inspired AI paradigms and hardware-software co-creation, to meet ambitious energy reduction goals and to secure the long-term profitability and sustainability of AI in networks.

Abbreviations

ACCEL	All–Analogue Chip Combining Electronics and Light
AI	Artificial intelligence
APIs	Application programming interfaces
CNNs	Convolutional neural networks
DNNs	Deep neural networks
EA4NG	Energy–efficient AI for Networks Guide
EU	European Union
ENVIAA	Environmental impact assessment of AI systems
FLOPS	Floating point operations per second
GPUs	Graphics processing units
HRM	Hierarchical reasoning models
HW	Hardware
IUCs	Internal use cases
LLMs	Large language models
LNNs	Liquid neural networks
ML	Machine learning
MoE	Mixture of experts
RNNs	Recurrent neural networks
SNNs	Sparse neural networks
SVMs	Support vector machines
TRM	Tiny recursive model

References

[1] R. Bellan, “Sam Altman says ChatGPT has hit 800M weekly active users,” TechCrunch, Oct. 06, 2025. Available: https://techcrunch.com/2025/10/06/sam-altman-says-chatgpt-has-hit-800m-weekly-active-users/

[2] S. Mehta, “How Much Energy Do LLMs Consume? Unveiling the Power Behind AI,” adasci.org, Jul. 03, 2024. Available: https://adasci.org/how-much-energy-do-llms-consume-unveiling-the-power-behind-ai/

[3] A. Silberling, “ChatGPT users send 2.5 billion prompts a day,” TechCrunch, Jul. 21, 2025. Available: https://techcrunch.com/2025/07/21/chatgpt-users-send-2-5-billion-prompts-a-day/

[4] S. Altman, “The Gentle Singularity,” Sam Altman Blog, Jun. 10, 2025. Available: https://blog.samaltman.com/the-gentle-singularity

[5] K. Alaamer, “This is the State of Play in the Global Data Centre Gold Rush,” World Economic Forum, Apr. 22, 2025. Available: https://www.weforum.org/stories/2025/04/data-centre-gold-rush-ai/

[6] A. Fleck, “Data Center Energy Consumption Surges Amid AI Boom,” Statista, Apr. 11, 2025. Available: https://www.statista.com/chart/34295/data-centers-electricity-generation-source/

[7] Nokia, “Beyond Speed: Sustainability with a 6G future-back design,” Nokia, Sep. 12, 2025. Available: https://www.nokia.com/6g/beyond-speed-sustainability-with-a-6g-future-back-design/

[8] TM Forum, “Autonomous Networks Framework v2.0.0 (IG1218F),” 9 May 2025. Available: https://www.tmforum.org/resources/introductory-guide/autonomous-networks-framework-v2-0-0-ig1218f/

[9] S. K. Moore, “The Secret to Nvidia’s AI Success,” IEEE Spectrum, Sep. 7, 2023. Available: https://spectrum.ieee.org/nvidia-gpu

[10] S. Pokhrel, “4 Popular Model Compression Techniques Explained,” xailient.com, Jan. 19, 2022. Available: https://xailient.com/blog/4-popular-model-compression-techniques-explained/

[11] B. Stevens, “Neural Magic CEO: We Shrink Machine Learning Models by 80-90%,” CDO Magazine, Dec. 11, 2024. Available: https://www.youtube.com/watch?v=rkN9OB5J_5s&t=23s

[12] CB Insights, “Best Compute-In-Memory (CIM) Companies,” CB Insights. Available: https://www.cbinsights.com/esp/enterprise-tech/data-management/compute-in-memory-(cim)

[13] K. Fitchard, “Bell Labs Prize Winners Have Designed a Computer Memory Chip for the Big-Data Era,” Nokia Blog. Available: https://www.nokia.com/blog/bell-labs-prize-winners-have-designed-a-computer-memory-chip-for-the-big-data-era/

[14] M. Murphy, “IBM Research’s Newest Prototype Chips Use Drastically Less Power to Solve AI Tasks,” IBM Research Blog, Aug. 23, 2023. Available: https://research.ibm.com/blog/analog-ai-chip-low-power

[15] Fortune Business Insights, “Future of AI with Human Brain-Like Functions: Numerous Opportunities Created by Key Neuromorphic Computing Companies,” Fortune Business Insights Blog, Aug. 28, 2024. Available: https://www.fortunebusinessinsights.com/blog/top-neuromorphic-computing-companies-11038/

[16] Y. Chen et al., “All-Analog Photoelectronic Chip for High-Speed Vision Tasks,” Nature, vol. 622, pp. 509–515, Oct. 25, 2023. Available: https://www.nature.com/articles/s41586-023-06558-8

[17] X. Luo et al., “Integer-Valued Training and Spike-Driven Inference Spiking Neural Network for High-Performance and Energy-Efficient Object Detection,” GitHub, Jul. 01, 2024. Available: https://github.com/BICLab/SpikeYOLO

[18] Y. Pan et al., “SpikingBrain: Spiking Brain-Inspired Large Models,” arXiv:2509.05276, Oct. 19, 2025. Available: https://arxiv.org/pdf/2509.05276

[19] Thousand Brains Project, “Reverse Engineering the Neocortex to Revolutionize AI.” Available: https://thousandbrains.org/

[20] A. Kosowski, “Baby Dragon Hatchling – Bridging the GAP Between Transformers and the Brain,” GitHub. Available: https://github.com/pathwaycom/bdh

[21] Liquid AI, “From Liquid Neural Networks to Liquid Foundation Models,” Liquid AI Research Blog, Sep. 30, 2024. Available: https://www.liquid.ai/research/liquid-neural-networks-research

[22] Sapient Intelligence, “Sapient Intelligence – We Are Building Self-Evolving Machine Intelligence to Solve the World’s Most Challenging Problems.” Available: https://www.sapient.inc/

[23] A. Jolicoeur-Martineau, “Less is More: Recursive Reasoning with Tiny Networks,” arXiv:2510.04871. Available: https://arxiv.org/abs/2510.04871

[24] K. Huang, H. Yin, H. Huang, and W. Gao, “Towards green ai in fine-tuning large language models via adaptive backpropagation,” arXiv preprint arXiv:2309.13192, 2023. Available: https://arxiv.org/abs/2309.13192

[25] S. Tuli, C.–H. Li, R. Sharma, and N. K. Jha, “CODEBench: A Neural Architecture and Hardware Accelerator Co-Design Framework,” ACM Trans. Embed. Comput. Syst., vol. 22, no. 3, May 2023. Available: https://doi.org/10.1145/3575798

[26] S. Negi et al., “Algorithm Hardware Co-Design for ADC-Less Compute In-Memory Accelerator,” IEEE Transactions on Circuits and Systems for Artificial Intelligence, vol. 1, no. 2, pp. 191–203, Dec. 2024. Available: https://ieeexplore.ieee.org/document/10750360

[27] AFNOR and Ecolab, “Référentiel général pour l’IA frugale,” The Shift Project, 2024. Available: https://greentechinnovation.fr/storage/2024/06/Referentiel-general-pour-lIA-frugale.pdf

[28] H. Smith and C. Adams, “Thinking about using ai?” Green Web Foundation, 2024. Available: https://www.thegreenwebfoundation.org/publications/report-ai-environmental-impact/

[29] S. Kallio and M. Farzan, “A Transparent and Methodical Framework to Assess the Sustainability Impact of AI,” Nokia, 2024. Available: https://onestore.nokia.com/asset/214115

[30] U. N. E. Programme, “Artificial intelligence (ai) end-to-end: The environmental impact of the full ai lifecycle needs to be comprehensively assessed – issue note,” 2024. Available: https://wedocs.unep.org/20.500.11822/46288

[31] AFNOR and Ecolab, “Référentiel général pour l’IA frugale,” The Shift Project, 2024. Available: https://tinyurl.com/4ef9mx7j

[32] D. Patterson et al., “Carbon emissions and large neural network training,” arXiv preprint arXiv:2104.10350, 2021. Available: https://arxiv.org/abs/2104.10350

[33] EcoLogits. Available: https://ecologits.ai/latest/

[34] GenAI Impact. Available: https://genai-impact.org/

[35] SUSTAIN–6G. Available: https://sustain-6g.eu/

NOKIA BELL LABS

Appendix: Reducing AI energy consumption in networks

How is energy used when developing AI solutions and what are the system level constraints?

Table 1. Energy consumption during AI lifecycle

High level	Step constraints	How it impacts energy	Network context examples
Data collection and processing	Collection [Compute, bandwidth, storage]	Sensors need to be powered while processing functions require energy to scrape or extract data on top of ordinary operation.	Data at different levels of the stack or signal processing chain such as antenna data for ISAC or operational data from event logs
	Retention [storage]	High sampling rate coupled with information rich data require large amounts of storage space and energy. Backups increase energy demand.	I/Q samples or FFT based data on a 5G frame basis (i.e., every 10ms) can generate many TBs of data.
	Preprocessing and cleaning [memory, compute, bandwidth, storage]	Labeling, feature extraction, normalization require significant processing and can be I/O intensive	Time domain (e.g., skew, variance) or frequency domain (e.g., DFT,MFCC) techniques to incorporate domain knowledge for feature presentation to AI model
Model selection and design	Design [memory, compute, bandwidth, storage]	Can require both high memory and compute	Task specific models (e.g., reinforcement learning for network congestion) are smaller in size while large models (e.g.,Nokia Language Model for understanding network terminology) are on the higher end of energy scale
	Learning efficiency [compute, storage, bandwidth]	A learning model that can continually learn will require less energy overall than having to retrain models from scratch	Most models used are off the shelf models and fine-tuned for the specific task and need to re-learn for new data.
Model implementation and deployment	Training [compute, bandwidth, memory size, data access]	Non-distributed training will require more hours while distributed training will require fewer hours, but energy will still be consumed across distributed paths in addition to synchronization overheads.	Most models for networks today use non-distributed training. Big AI Models for wireless networks with collaborative edge computing are being researched.
	Inference [compute, bandwidth, memory size, data access]	For inference that requires access to large data stores e.g., RAGs increase energy needs.	Nokia Language model uses a data store that stores network terminology in vector form and is used to search through a trove of technical documents.
	Data movement [bandwidth]	Moving data to a processing location requires additional bandwidth and energy	Critical tasks such as adversarial object detection or safety monitoring using wireless signals may require multiple GB/s transfer rate for data to reach AI models

What are the system level constraints to consider?

Table 2: Constraints to consider during AI solution development

Constraints	How it impacts the AI solution
Data access (e.g., traditional von Neumann versus "near or in-compute" memory)	• Bottlenecks in data access significant contributor to low AI processor utilization • New compilers/framework required to take advantage of new compute architecture
Memory size	• CPU: needed for loading datasets and model weights, storing feature vectors before transferring to GPU memory. Also limits in memory data sizes for non-AI related tasks. • GPU: limits model weight size and data batch size
Number of cores per node/machine	• CPU: for parallel processing and handling multiple data streams • GPU: determines model parallel processing capacity
Interconnect bandwidth	• Impacts how fast data can move – GPU waiting for data is a common bottleneck which reduce utilization
Latency	• Determines how fast information can be derived for downstream tasks
Storage	• While cost per GB is low, it is not infinite

The following table shows a pragmatic approach to considering all of the constraints listed above.

Table 3. Guidelines for energy reduction

High-level step	Action	Description	Energy impact
Data collection and processing	Engineer the training data using subject matter expertise	Model performance and training times can be reduced if the training data is structured, filtered and has a feature set matched to the task.	Besides, performance boost, reduces energy associated with data movement, storage and training.
Model selection and design	Decide on learning model architecture: classical versus modern	Traditional machine learning models can be effective and consume less energy. Modern deep learning models are more flexible but can consume relatively significant amount of energy.	The selection of classical AI results in reduced energy for computation and on-board memory support creating space for other models/applications.
	Use domain knowledge	Stay close to classical signal processing and use AI where it makes sense. Don't replace functions where a classical algorithm works well. Do replace functions that are heuristic, tuned by experimentation or are data dependent.	Strongly reduced model sizes mean proportional savings in energy. Starts from classical baseline performance.
	Slim down neural network architecture	Use known methods to reduce model size (depth-wise separable convolutions, grouped convolutions, ...)	Further reduction in the number of model parameters and compute and therefore energy savings.

High-level step	Action	Description	Energy impact
	HW/SW co-design and quantization	Quantization of weights and activations can reduce compute further, but only if the quantized operations can be mapped to hardware. Quantization aware training is necessary for best performance and different parts of the model can have different precision requirements.	Model that is closely linked to hardware and makes use of the most energy efficient number representations. Smaller precision on the right hardware translates to energy savings in elementary operations (add, multiply).
Model implementation and deployment	Use right amount of flexibility	Less flexible AI models can be made more energy efficient (deterministic data movement) but sacrifice usability. This trade-off needs to be carefully considered for deployment.	Hardcoding of most compute-intense parts of a model reduces data transfer and therefore energy usage.
	Shrink models for particular use cases	If the application allows, split it into easy and hard use cases. Use smaller models for easy use cases.	Use case specific models for average reduction in computational effort/energy.

Nokia OYJ
Karakaari 7
02610 Espoo
Finland
Tel. +358 (0) 10 44 88 000

Document code: SR1864550EN (December) CID215230