E2E energy monitoring for AI inference in mobile networks

Abhishek Dandekar
a.dandekar@tu-berlin.de
TU Berlin
Berlin, Germany

Ashrafur Rahman
a.rahman@tu-berlin.de
TU Berlin
Berlin, Germany

Julius Schulz-Zander
julius.schulz-zander@hhi.fraunhofer.de
Fraunhofer HHI
Berlin, Germany

Abstract

The increasing adoption of Artificial Intelligence (AI), particularly large language (LLMs) and vision-language models (VLMs) has led to a sharp rise in energy demand. While most of the studies predominantly assess energy consumption during training and inference, they often neglect the energy required to transport contextual data—such as text, images, or video—from far-edge devices to AI models, especially over mobile networks. We measure and analyze energy consumption for AI inference both on model-level and network-level. Our approach leverages a combined cross-layer and in-band network telemetry approach to estimate application-level energy usage. Our experiments show that the energy used by the network can be on par with that used by energy efficient AI models for certain tasks. Furthermore, we also estimate the total CO_2 emissions of these inference workflows. These results highlight the critical need to incorporate network consumed energy into sustainable AI system design.

CCS Concepts

• **Networks → Network performance analysis**; *Network measurement*.

Keywords

Energy, Green 6G, Carbon aware networks, AI inference, Telemetry

ACM Reference Format:
Abhishek Dandekar, Ashrafur Rahman, and Julius Schulz-Zander. 2025. E2E energy monitoring for AI inference in mobile networks. In *1st Workshop on Next-Generation Network Observability (NGNO '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3748496.3748995

NGNO '25, September 8-11, 2025, Coimbra, Portugal

ACM ISBN 979-8-4007-2087-1/2025/09
https://doi.org/10.1145/3748496.3748995

1 Introduction

The use of AI, particularly LLMs, has skyrocketed in recent years. This in turn has driven up demand for compute and thereby more energy. It is projected that the electricity consumption due to AI will more than double to around 945 TWh by 2030[4]. This is primarily due to the increasing sizes of the AI models and also increasing usage of AI across sectors. AI energy consumption majorly happens in two stages, during training and during inference. For the scope of this paper, we focus on energy consumed during inference.

LLM inferencing consists of two key inputs, the prompt which describes the instructions from the user, and the relevant context. The relevant context might consist of the latest information that was not present at the time of model training. This allows the AI model to work with the latest information. The context is generally produced by far-edge entities like cars, robots, etc., and has to be transported to the AI model. In many use cases like autonomous cars, robots in smart factories, this context is carried over the mobile network. The context can be in the form of multiple modalities, like text, images, videos, etc. AI energy monitoring frameworks generally only consider energy consumed by the AI model during inference and ignore energy consumed by transporting the context from data source to the AI model. In this paper, we investigate the energy consumed by the AI models, specifically for VLMs, along with the energy consumed by transport of context over mobile network.

In section 2, we provide relevant background. In section 3, we provide overview of related literature. In section 4, we provide overview of the energy monitoring framework. In section 5, we present and analyze the results. In section 6, we provide the limitations of the work, and in section 7, we present the conclusion.

2 Background

2.1 Vision Language Models

Vision-Language Models (VLMs) are a major step forward in multimodal artificial intelligence. They combine computer vision and natural language processing, enabling them to understand and reason over both images and text. As a result, VLMs can perform a wide range of tasks, such as image captioning, visual question answering (VQA), image-text retrieval, and multimodal content generation. Unlike earlier computer vision models that were limited to specific tasks, VLMs are designed to be more general and adaptable. They are trained on large-scale datasets containing paired images and text, and often use transformer-based architectures. This allows them to generalize to new tasks with little or no additional training, making them effective across various domains and use cases.

Despite their exceptional performance, VLMs require significant computing resources and thereby energy. These models are large, often with billions of parameters, and usually require servers with dedicated Graphical Processing Units (GPUs). They are usually located in cloud or edge datacenters and depend on computer or mobile networks to fetch relevant context for inference.

2.2 5G network

3GPP 5G[3] is a mobile network technology that allows devices to send data at higher speeds and lower latency over the air. It consists of three major parts, the Radio Access Network (RAN), Core Network (CN), and the Transport Network (TN). These networks consist of Virtual Network Functions (VNFs) and Physical Network Functions (PNF). VNFs are software components in container running on a general purpose compute servers. PNFs are hardware components generally designed to perform specific narrow functions. VNFs and PNFs work together to form the 5G network, which allows transporting data from user equipment (UE) to its destination.

O-RAN[18] is a set of specifications that build on top of 3GPP 5G and make the RAN more flexible. In O-RAN's disaggregated architecture, the RAN consists of centralized unit (CU), distributed unit (DU) and radio unit (RU). The CU handles the higher layer protocols in the RAN while the DU executes lower layer protocol functions. The RU is responsible for the actual transmission and reception of radio signals. All these RAN components are controlled using two radio intelligent controllers (RICs). The non-real time RIC handles coarse-grained parameters on a wider scale of the network, while the near-real time RIC controls RAN parameters using different services models.

3 Related Work

Energy and carbon monitoring for AI inference workloads is being studied widely due to the recent upsurge in AI usage. In [12], Desislavov et al. analyze trends in inference energy consumption and find that, despite growing model sizes, improvements in hardware have not caught up which affects their energy footprint. Li et al. [16] propose *EcoServe*, a carbon-conscious orchestration system designed for large-scale LLM inference in cloud environments. Similarly, in *Clover*[15] authors propose a carbon-aware inference runtime that prioritizes emissions reduction over traditional performance metrics.

Monitoring networks for energy efficiency and carbon emissions is also being investigated across academia and industry. In [20] authors present the mobile network operators perspective on roadmap to Green 6G networks. They explore key technologies that could enable such networks. In [10], authors describe challenges & opportunities at different network levels to increase energy efficiency. In [14] authors describe how monitoring mobile network carbon emissions at lower granularity can enable new kinds of services for future networks. In [11], a granular energy monitoring framework is developed, however the measurements are done on a emulated network without a real radio. In [8], the authors demonstrate how carbon metrics can be monitored in transport networks using telemetry aggregation.

4 E2E energy monitoring framework

Consider a scenario where a certain process is being monitored by a robot using a camera in a smart factory environment. A video stream from the camera is sent to a VLM running on a server in the edge or cloud for further analysis. In order to monitor end-to-end energy consumption in such an environment, there are two challenges. Firstly, the robot might be running multiple applications which use the 5G network like command and control, telemetry, etc. Each of these applications might have various QoS requirements, hence it might be routed through different NFs (e.g, UPF). These NFs might have varying energy consumption. Hence, it is challenging to identify which application consumes what amount of energy in the network. Secondly, as most of the NFs in 5G are virtualized, it makes it harder to measure energy consumption as compared to PNFs. We create a monitoring framework to address both of these challenges as seen in Fig. 1.

4.1 Application data path tracing

In order to trace the datapath taken by the video streaming application over the 5G link, this framework uses In-Situ Operations, Administration, and Maintenance (IOAM) standard[9]. IOAM allows sending operational and telemetry information directly within the data packets as they traverse a network path, rather than using separate, dedicated telemetry packets. We implement this by embedding IOAM header in IPv6 packet header[13]. Each packet of the video stream traffic carries this IOAM header. We use two fields in this header, *TraceID* and *NodeID*.

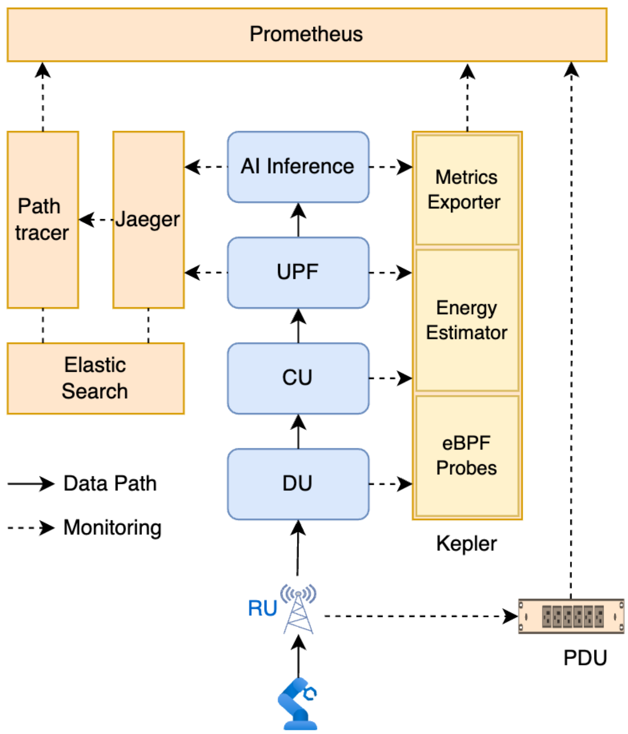

Figure 1: E2E monitoring framework

- *TraceID*: Unique ID identifying the application to which the packets belong. Inserted at the source of the data traffic.
- *NodeID*: This uses the Hop-by-Hop Options header feature of IOAM. As the packet passes through each of the VNFs in the datapath, each VNF adds its unique *NodeID* to the header. When the traffic is received at the AI inference server, it has a list of NodeIDs for all the VNFs the packet passed through.

As seen in Fig. 1, IOAM telemetry data is extracted from the user data packets and is collected by Jaeger, which is a distributed tracing platform. This data is then analyzed by the path tracer. By correlating *NodeID* and *TraceID* from the received data, the path tracer determines the number of packets from a specific application traversing each set of NFs[11]. These traces can be stored in Elasticsearch for future use.

4.2 Energy measurements

An O-RAN compliant 5G network consists of multiple NFs in the datapath of the video stream traffic. The RU is the only PNF in the datapath, while the rest are VNFs. We monitor the energy consumption of RU using a metered power distribution unit (PDU). These measurements are then pushed to Prometheus which is an event monitoring framework. For measuring energy consumed by VNFs like DU, CU, UPF, we use Kepler[6]. Kepler allows monitoring the energy consumption of each pod in a Kubernetes cluster. It achieves this by collecting and fusing data from multiple sources like running average power limit (RAPL), advanced configuration and power management interface (ACPI), and Berkeley Packet Filter (BPF) based data capture of hardware performance counters. We also use Kepler to measure the energy consumed by AI model inference. The energy consumption data for each of the NFs and the AI model is then forwarded by Kepler to Prometheus for further analysis, as seen in Fig. 1.

4.3 Application level energy estimation

Energy measurement of each NF and AI model has two components- static and dynamic. Static component is energy that is being consumed irrespective of the load, while the dynamic component is the one that changes according to changes in load. Based on these, we can classify the NFs into three types:

- Zero Static energy consumption: These functions only have a dynamic component, i.e., they consume energy only when loaded (e.g., AI model inference)
- Constant static energy consumption: In these functions, static energy consumption is constant while the dynamic component is negligible. They keep consuming energy irrespective of load (e.g., RU).
- Constant static and variable dynamic energy consumption: These functions consume a constant amount of energy irrespective of load, and the dynamic energy consumption changes according to load (e.g., CU).

In order to calculate the energy consumed by an application in a given network function we use the following formula.

$$E(App)_{NF} = \left(\frac{Pkt(App)_{NF}}{Pkt(Total)_{NF}} \times E(NF)^{dyn} \right) + \left(\frac{E(NF)^{static}}{MaxCap_{NF}} \times Rate(App) \right)$$

We take the proportion of application packets passed through an NF to the total number of packets passing through the NF for a given amount of period and multiply it by the dynamic energy component of the NF. As some NFs have large static components, in order to have fair attribution, we divide the static component by the maximum possible capacity for the NF and multiply it by application data rate. The addition of these two parts gives us energy used by the application in a given network function- $E(App)_{NF}$. In order to get the end-to-end energy consumption, we add $E(App)_{NF}$ for each network function, and then add the dynamic energy component of the AI model inferencing. Note that there is no static energy component for AI model inferencing.

$$E(App)_{E2E} = E(App)_{RU} + E(App)_{DU} + E(App)_{CU} + E(App)_{UPF} + E(App)_{AI}^{dyn}$$

5 Results and Analysis

5.1 Testbed overview

In order to measure realistic E2E energy measurement, we set up a testbed where a UE transmits video over a real private 5G network (4Mbps50 fps). We chose this rate as it represents datarate typically required by monitoring cameras. The UE monitors an assembly line where a set of boxes are moving along a conveyor belt. This video is received by a VLM model, which analyses the video according to a pre-defined prompt. Note that this can be a typical scenario in a smart factory environment with a private 5G network[5]. The AI model samples the video at a rate of 1.6fps. This rate was chosen as it is the typical VLM sampling rate for most applications, however for special applications, a higher or variable sampling rate might be necessary[19]. We run inference using the sampled video frames and prompt every 5 seconds. Each experiment is performed for 60 minutes. We use two VLMs, Qwen2.5 VL[7] and gemma3[22] for evaluation. Both models run on a NVIDIA A100 GPU with 80GB VRAM.

Our O-RAN compliant private 5G network consists of a Liteon FlexiFi RU, the DU and CU from OpenAirinterface[1], and the core network from Open5GS[2]. The 5G VNFs run on a server with 32 Intel Xeon Gold 6346 CPUs with 512GB RAM. The RU is connected to the DU via an optical fiber through a fronthaul switch. Both 5G VNFs and AI models run as Kubernetes pods.

In order to evaluate the impact of different kinds of AI tasks on energy consumption, we classify tasks into two categories- simple and complex. Both tasks use the same video stream as context but different prompts. In simple tasks, the VLM is tasked with identifying the object in the video or answering yes/no questions about the objects in the video. In complex tasks, the VLMs is tasked with more analytical or descriptive tasks like describing what's happening in the video or predicting what might happen next. In simple tasks, the VLM output consists of 1 or 2 words, while in complex tasks, the output is much more descriptive. VLM tends to consume more energy in complex tasks than in simpler tasks, especially as the number of output tokens rises[17].

5.2 Energy consumption

We compare energy consumed by the 5G network with energy consumed by the gemma3 and Qwen2.5VL models for simple tasks as seen in Fig. 2. The video stream is being sent at rate of 4 Mbps. We find that gemma3 models consume approximately 4x more energy than Qwen2.5 models. The more unexpected finding is that 5G network spends approximately same amount of energy to transport the video context as Qwen2.5 spends in inferencing.

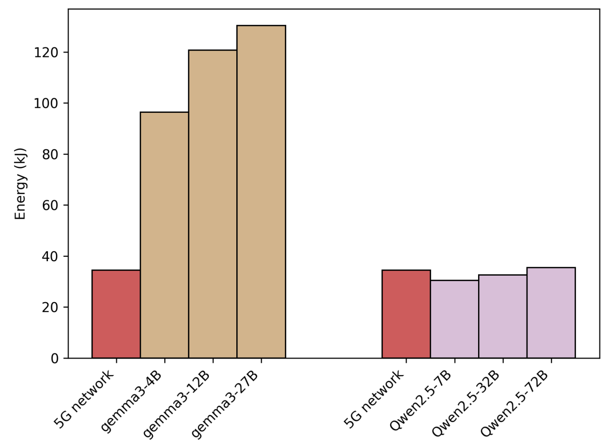

Figure 2: Energy consumption for simple queries

Similarly, in Fig. 3, we compare energy consumed by the 5G network with energy consumed by VLMs for complex tasks. In this case, with the exception of the Qwen2.5-7B, both gemma3 and Qwen2.5 models exhibit similar energy consumption. For this scenario energy consumed by the network is 20% to 30% of the energy consumed by larger VLMs.

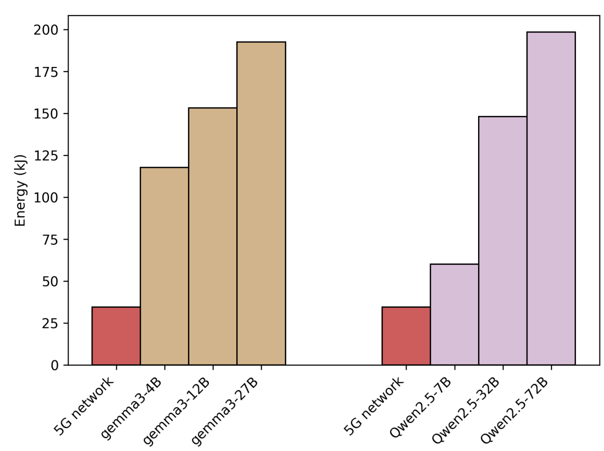

Figure 3: Energy consumption for complex queries

We also find that sending the context at a slower rate reduces the energy consumption. We try to examine if selecting the more energy efficient model for simple queries and sending the context at a lower rate are together able to lower the overall energy consumption. In Fig. 4 the left group of bars shows total energy consumption when video is sent at 4 Mbps and gemma models are used. Bars on the right show total energy consumption when Qwen2.5 models are used and video is sent at 2 Mbps.

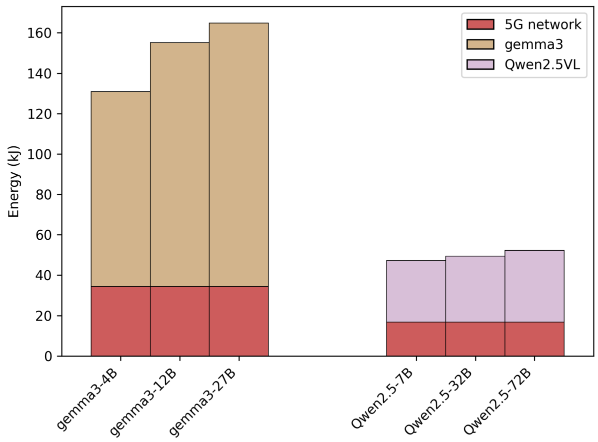

Figure 4: Energy saving impact: Simple tasks

We find that choosing the energy efficient AI model and lower rate might reduce energy consumption by 68%. We also examine if a similar reduction is also possible with complex tasks, as shown in Fig. 5. However, the energy reduction here is very negligible (0.1%) with the exception of smaller models. Note that sending video stream at lower rates might only be feasible for certain kinds of use cases(e.g., retrospective analysis) and may not suit ones with more strict timing requirements.

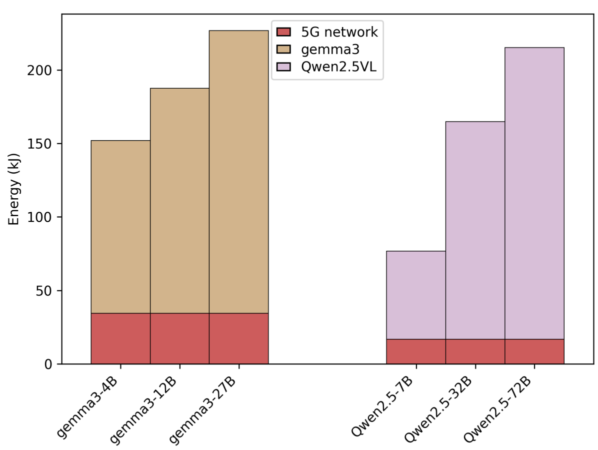

Figure 5: Energy saving impact: Complex tasks

Also note that both model families have comparable accuracy, and it can be further improved by fine-tuning.

5.3 Carbon intensity

The environmental impact of energy consumption is determined by the carbon intensity. Carbon intensity is the amount of CO_2 emitted per kWh of energy consumption. The volume of CO_2 emissions varies depending on whether the energy source is fully renewable or hybrid (mix of renewable and non-renewable energy sources). While significant attention has been directed toward reducing the carbon emissions associated with AI model inference, comparatively less emphasis has been placed on the emissions resulting from network infrastructure during data transmission.

It is relatively easier for data centers to switch to a renewable energy source due to their centralized nature. In case of mobile network, due to its distributed nature, it might be relatively difficult to operate on a fully renewable energy source. In several developing countries, mobile network operators use diesel generators to power mobile base stations [21].

We calculate total carbon emitted for simple and complex tasks as shown in Fig. 6 and 7. We assume the data center running VLM is powered by fully renewable energy (25g CO_2/kWh) while the mobile network is using a hybrid mix (215g CO_2/kWh)[23]. Under these conditions, network-induced carbon emissions may surpass those generated by the AI inference process. We find that for simpler tasks, the network is responsible for around 70% to 89% of total emissions, depending on the model used. In case of complex tasks, the network is responsible for approximately 60% to 84% of the total emissions. This highlights the need of focusing not only on energy consumption but also on the associated carbon emissions.

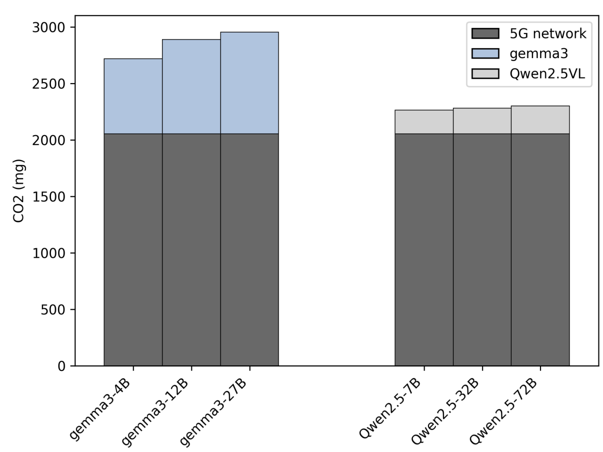

Figure 6: E2E CO_2 emissions: Simple tasks

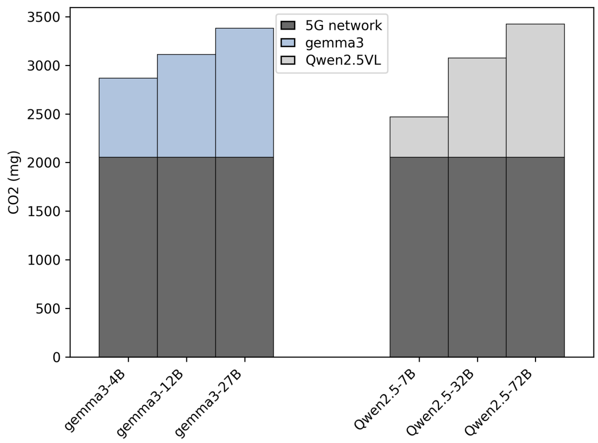

Figure 7: E2E CO_2 emissions: Complex tasks

6 Limitations

- We assume that the dynamic energy component of each NF scales linearly with the number of packets processed. This may not be true for some NFs. In such cases, these NFs need to be benchmarked and an appropriate scaling factor needs to be applied to the measured energy.
- Our work focuses only on data plane energy consumption and does not take control plane energy consumption into account. Control plane consumption is estimated to be approximately 10% of the overall data plane consumption.
- Adding an IOAM header to a packet adds around 11.83% of overhead on the packet in terms of size. It also contributes to around 13% energy overhead. However, it is possible to enable IOAM mode only when actively measuring energy and turning it off when it is not required to save overheads.
- We do not consider embodied emissions like emissions during VLM training, manufacturing, and assembly of network components, etc. We only focus on operational emissions.

7 Conclusion and Future work

In this paper, we present a framework to monitor end-to-end energy consumption for AI inference over mobile networks. Our framework uses a combined cross-layer and in-band network telemetry approach to estimate application-level energy usage. Our findings reveal that for simpler tasks executed with energy-efficient VLMs, the energy consumed by the mobile network is approximately on par with the energy required for AI inference by the VLM. In contrast, for more complex tasks, network energy consumption constitutes roughly 20% to 30% of the total energy used by the VLM. Moreover, we demonstrate that combining energy-efficient AI models with lower data transmission rates can reduce overall energy consumption by up to 68%. These findings highlight the pressing need for advancements in the energy efficiency of mobile networks, which must progress in tandem with improvements in AI model and hardware efficiency to enable end-to-end sustainable AI systems.

In future we aim to benchmark energy consumption for more VLM models and with more number of UEs and traffic. We also aim to optimize the IOAM header to reduce the overhead.

This work does not raise any ethical issues.

Acknowledgments

The authors thank the Q-net-Q Project which has received funding from the European Union's Digital Europe Programme under grant agreement No 101091732, and is cofunded by the German Federal Ministry of Education and Research (BMBF). The authors also acknowledge the financial support by the BMBF in the program of "Souverän. Digital. Vernetzt." Joint project 6G-RIC (16KISK020K).

References

[1] 2024. oai / openairinterface5G. https://gitlab.eurecom.fr/oai/openairinterface5g/-/tree/develop. Accessed: 2024-07-13.
[2] 2024. open5gs. https://github.com/open5gs/open5gs. Accessed: 2024-07-13.
[3] 2025. *3GPP TS 23.501 v19.3.0.* https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144
[4] 2025. *Energy and AI – Analysis - IEA.* https://www.iea.org/reports/energy-and-ai
[5] 2025. Fujitsu Verifies Effectiveness of Private 5G in Manufacturing Sites in Collaboration with Microsoft Japan - Fujitsu Global. https://www.fujitsu.com/global/about/resources/news/press-releases/2020/1008-02.html. Accessed: 2025-07-14.
[6] Marcelo Amaral, Huamin Chen, Tatsuhiro Chiba, Rina Nakazawa, Sunyanan Choochotkaew, Eun Kyung Lee, and Tamar Eilam. 2023. Kepler: A Framework to Calculate the Energy Consumption of Containerized Applications. In *2023 IEEE 16th International Conference on Cloud Computing (CLOUD).* 69–71. doi:10.1109/CLOUD60044.2023.00017
[7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and OTHERS. 2025. Qwen2.5-VL Technical Report. (2025). arXiv:2502.13923 [astro-ph.IM] http://arxiv.org/pdf/2502.13923
[8] Ramon Bister, Alexander Clemm, Severin Dellsperger, and Reto Furrer. 2024. Towards Sustainable Networking: Unveiling Energy Efficiency Through Hop and Path Efficiency Indicators in Computer Networks. In *2024 IEEE 10th International Conference on Network Softwarization (NetSoft).* 118–126. doi:10.1109/NetSoft60951.2024.10588907
[9] F. Brockners, S. Bhandari, and T. Mizrahi. 2022. RFC 9197: Data Fields for In Situ Operations, Administration, and Maintenance (IOAM).
[10] Alexander Clemm and Cedric Westphal. 2022. Challenges and Opportunities in Green Networking. In *2022 IEEE 8th International Conference on Network Softwarization (NetSoft).* 43–48. doi:10.1109/NetSoft54395.2022.9844020

[11] Abhishek Dandekar, Johannes Wegener, Ashrafur Rahman, and Julius Schulz-Zander. 2024. Towards Application Level Energy Monitoring for Green 6G Networks. In *Proceedings of the 4th ACM Workshop on 5G and Beyond Network Measurements, Modeling, and Use Cases* (Los Angeles, CA, USA) *(5G-MeMU '24)*. Association for Computing Machinery, New York, NY, USA, 8–13. doi:10.1145/3694810.3700158

[12] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems* 38 (2023), 100857.

[13] Justin Iurman, Frank Brockners, and Benoit Donnet. 2021. Towards cross-layer telemetry. In *Proceedings of the Applied Networking Research Workshop (ANRW '21)*. ACM, 15–21. doi:10.1145/3472305.3472313

[14] Rashmi Kamran, Shwetha Kiran, Pranav Jha, Abhay Karandikar'T, and Prasanna Chaporkar. 2024. Green 6G: Energy Awareness in Design. In *2024 16th International Conference on COMmunication Systems & NETworkS (COMSNETS)*. 1122–1125. doi:10.1109/COMSNETS59351.2024.10427334

[15] Baolin Li, Siddharth Samsi, Vijay Gadepally, and Devesh Tiwari. 2023. Clover: Toward Sustainable AI with Carbon-Aware Machine Learning Inference Service. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '23)*. ACM, 1–15. doi:10.1145/3581784.3607034

[16] Yueying Li, Zhanqiu Hu, Esha Choukse, Rodrigo Fonseca, G. Edward Suh, and Udit Gupta. 2025. EcoServe: Designing Carbon-Aware AI Inference Systems. arXiv:2502.05043 [cs.DC] https://arxiv.org/abs/2502.05043

[17] Soham Poddar, Paramita Koley, Janardan Misra, Sanjay Podder, Niloy Ganguly, and Saptarshi Ghosh. 2025. Towards Sustainable NLP: Insights from Benchmarking Inference Energy in Large Language Models. (2025). arXiv:2502.05610 [astro-ph.IM] http://arxiv.org/pdf/2502.05610

[18] Michele Polese, Leonardo Bonati, Salvatore D'Oro, Stefano Basagni, and Tommaso Melodia. 2023. Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges. *IEEE Communications Surveys & Tutorials* (2023), 1376–1411. doi:10.1109/COMST.2023.3239220

[19] Tianyuan Qu, Longxiang Tang, Bohao Peng, Senqiao Yang, Bei Yu, and Jiaya Jia. 2025. Does Your Vision-Language Model Get Lost in the Long Video Sampling Dilemma? (2025). arXiv:2503.12496 [astro-ph.IM] http://arxiv.org/pdf/2503.12496

[20] Antonio De Domenico Rishikesh Chakraborty. 2024. Green future networks: A Roadmap to Energy Efficient Mobile Networks - NGMN. https://www.ngmn.org/publications/green-future-networks-a-roadmap-to-energy-efficient-mobile-networks-2.html. Accessed: 2024-07-13.

[21] sjena@gsma.com. 2025. Why and how mobile operators are looking to renewables to power networks across Africa | Mobile for Development. https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-for-development/blog/why-and-how-mobile-operators-are-looking-to-renewables-to-power-networks-across-africa/. Accessed: 2025-07-14.

[22] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and OTHERS. 2025. Gemma 3 Technical Report. (2025). arXiv:2503.19786 [astro-ph.IM] http://arxiv.org/pdf/2503.19786

[23] Vattenfall. 2024. Vattenfall Stromkennzeichnung – Transparenz in den Tarifen. https://www.vattenfall.de/stromkennzeichnung. Accessed: 2024-07-14.