Ph.D. proposal:

FAST — Fast Awareness Raising of AI Services Energy Impact

Ludovic APVRILLE, Bastien SULTAN
Telecom Paris
LabSoC team
450 routes des Chappes, F-06904 Sophia-Antipolis Cedex, France
Email: firstname.lastname@telecom-paris.fr

July 10, 2026

1 Context and research problem

Artificial intelligence functions are increasingly integrated into embedded and cyber-physical systems used in transportation, industrial production, telecommunications, and biomedical applications. Examples include perception and decision support in vehicles, predictive maintenance, medical decision support, railway supervision, and adaptive communication systems. In such systems, energy consumption does not result solely from training an AI model. It also depends on the complete operational chain: sensing, data transfer, preprocessing, memory accesses, inference, hardware acceleration, activation policies, real-time constraints, and the frequency and duration of use.

The environmental impact of AI has become an important research topic. Existing work has highlighted the need to report computational resources and energy use as first-class evaluation criteria [9, 5]. Other studies provide methods for estimating the energy or carbon footprint of machine-learning workloads [10, 6, 2]. Research on efficient neural-network execution and model compression has also produced significant advances [12, 4]. Nevertheless, most available approaches rely on measurements performed after implementation, focus on cloud infrastructures or model training, or require technical expertise and detailed execution data.

Embedded AI creates additional difficulties. Energy consumption depends strongly on the target hardware, the distribution of computation between edge and cloud resources, communication technologies, timing constraints, data rates, and operational scenarios. Designers therefore lack a lightweight method for estimating energy use during the early design stages, when major architectural and functional decisions can still be changed. Service providers and users also have limited awareness of the environmental consequences of activating or repeatedly using AI-enabled functions.

The FAST project addresses this gap by combining system-level modelling and simulation with a field study of organisational practices. The central research problem is the following:

> How can the energy impact of embedded AI services be estimated rapidly from high-level system and usage models, and how can this information support more environmentally responsible design and usage decisions?

This problem is both technical and socio-organisational. A technically accurate estimator is insufficient if its results are difficult to understand, are delivered too late, or do not correspond to the decisions made by system architects, service providers, and end users. Conversely, awareness-raising mechanisms must rely on credible estimates and explicitly represent uncertainty.

2 Objectives

The main objective of this Ph.D. thesis is to define, implement, and evaluate a modelling and simulation method that rapidly estimates the energy impact of embedded AI services without requiring a complete implementation or physical execution of the target system.

The thesis will pursue four complementary objectives:

1. Define representative usage scenarios for embedded AI systems in one or two application sectors, selected among healthcare, automotive, and railway systems.
2. Develop a high-level modelling formalism describing AI functions, embedded hardware and software resources, communications, deployment choices, and usage parameters.
3. Implement a simulation and estimation tool that provides energy indicators, sensitivity analyses, comparisons between design alternatives, and explicit uncertainty ranges.
4. Evaluate how access to these indicators changes design choices, usage practices, organisational routines, and awareness of environmental impact.

The resulting approach should support both designers and users. For designers, it should help compare architectures, deployment strategies, AI models, accelerators, communication patterns, and activation policies. For service providers and users, it should make the energy consequences of usage scenarios understandable and actionable.

3 Scientific foundations and related work

The technical part of the thesis will build on model-based systems engineering, embedded-system simulation, and architecture exploration. Previous work by the host team has investigated automated enrichment of SysML models through mutation [11] and AI-assisted system modelling [1]. These results provide relevant foundations for extending system models with energy-related information and for reducing the effort required to construct or enrich models.

At the AI level, the thesis will draw on Green AI principles [9], systematic reporting of energy and carbon footprints [5], and energy-aware execution of deep neural networks [12]. Compression, quantisation, pruning, conditional execution, and hardware acceleration will be represented as design alternatives rather than treated only as low-level implementation optimisations [4].

The socio-organisational part will analyse how environmental information is interpreted and integrated into collective activity. Awareness is not merely the display of information; it depends on the context of work, the actors' responsibilities, and the way information supports coordination and action [8]. Organisational routines may be reconfigured in response to crises or ecological challenges [7, 3]. These perspectives will guide the design of indicators and the longitudinal evaluation of their effects.

4 Expected work and methodology

The research will follow an iterative, interdisciplinary methodology combining model construction, software implementation, experimental calibration, and field studies.

1. **State of the art and requirements elicitation.** Review methods for energy modelling of embedded systems, AI inference, edge/cloud allocation, communication subsystems, and environmental assessment. Conduct interviews and observations with domain experts to identify decisions, constraints, representative scenarios, and appropriate indicators.

2. **Usage-scenario definition.** Formalise representative workflows for one or two sectors. A scenario will specify actors, AI functions, input data, execution frequency, duration, quality-of-service constraints, hardware targets, communication paths, and possible alternatives.

3. **Energy-aware system metamodel.** Define modelling concepts for software tasks, AI operators or models, processors, accelerators, memories, sensors, networks, power states, deployment allocations, and usage parameters. The metamodel should support hierarchical refinement and reusable component profiles.

4. **Estimation and simulation engine.** Develop a compositional estimation method combining parameterised energy models with discrete-event or scenario-based simulation. The engine will calculate energy per request, per operational cycle, and over a specified period. It will also identify dominant contributors and compare alternatives.

5. **Technical validation.** Compare simulated results with physical measurements using power monitors or on-board sensors. Evaluate accuracy, execution time, robustness, and portability across platforms and scenarios.

6. **Field evaluation.** Study practices and environmental awareness before and after introduction of the tool. Data may include interviews, questionnaires, observations, design artefacts, tool traces, and documented decisions. The analysis will examine whether the tool changes problem framing, trade-offs, routines, and actual usage.

5 Indicative schedule

Period	Main activities
Months 1–6	Literature review, selection of application fields, field-study protocol, initial interviews, scenario and requirement definition.
Months 7–12	Energy-aware metamodel, parameter taxonomy, first reusable profiles, prototype architecture.
Months 13–18	Simulation-engine implementation, demonstrator instrumentation, initial calibration.
Months 19–30	Design-space exploration, multi-scenario technical validation.
Months 25–30	Participatory design of feedback mechanisms, pilot deployment, before/after field study.
Months 31–36	Consolidation and release of the prototype, publications, dissertation writing and defence.

6 Scientific environment

The project combines expertise in embedded systems, system modelling, verification, artificial intelligence, and organisational studies. Potential contributors within Institut Mines-Télécom include Ludovic Apvrille and Bastien Sultan at Telecom Paris, and Philippe Jaillon at Mines Saint-Étienne. Loubna Echajari at the University of Technology of Troyes may contribute expertise in organisational routines and ecological transitions.

Previous collaborations include the European SPARTA project, PERP5G, and the CMA TCE programme. Ludovic Apvrille, Bastien Sultan, and Loubna Echajari have also collaborated on the preparation of a joint French National Research Agency proposal.

The broader FAST initiative may ultimately involve several complementary Ph.D. projects addressing energy modelling and simulation, frugal AI and optimisation, and organisational practices. The present thesis is nevertheless defined as an autonomous project centred on the modelling tool and its sociotechnical evaluation.

7 Required skills

- Strong background in computer science, computer engineering, embedded systems, systems engineering, or software engineering.
- Experience in modelling, simulation, optimisation, machine learning, or embedded AI is desirable.
- Programming skills and an interest in experimental evaluation are required.
- Interest in interdisciplinary research, field studies, and environmental issues related to digital technologies.
- Good scientific English. Knowledge of French would facilitate interaction with French-speaking field-study participants.

8 How to apply

Applications should be sent as a single PDF file to ludovic.apvrille@telecom-paris.fr. The application should contain:

- a curriculum vitae;
- a cover letter explaining the candidate's interest in the topic and research environment;
- Master's degree transcripts;
- recommendation letters or contact details for academic referees;
- when available, examples of research, software, reports, or publications relevant to the topic.

Candidates will be assessed on their technical background, their ability to conduct scientific research, and their motivation for interdisciplinary work.

References

[1] Ludovic Apvrille and Bastien Sultan. System architects are not alone anymore: Automatic system modeling with AI. In Francisco José Domínguez Mayo, Luís Ferreira Pires, and Edwin Seidewitz, editors, *Proceedings of the 12th International Conference on Model-Based Software and Systems Engineering (MODELSWARD 2024)*, pages 27–38, Rome, Italy, February 2024. SCITEPRESS. Conference held 21–23 February 2024.

[2] Jesse Dodge, Taylor Prewitt, Rémi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A. Smith, Nicole DeCario, and Will Buchanan. Measuring the carbon intensity of AI in cloud instances. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM International Conference Proceeding Series, pages 1877–1894, New York, NY, USA, June 2022. Association for Computing Machinery. FAccT '22, Seoul, Republic of Korea, 21–24 June 2022.

[3] Loubna Echajari, Nathalie Lazaric, and Dorota Leszczyńska. Reconfiguring practices and routines to meet ecological challenges: in vino veritas. Paper presented at the 36th EGOS Colloquium, "Organizing for a Sustainable Future: Responsibility, Renewal & Resistance," held virtually and hosted by the University of Hamburg, 2–4 July 2020, July 2020.

[4] Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings*, San Juan, Puerto Rico, 2016.

[5] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 21(248):1–43, 2020.

[6] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning, October 2019.

[7] Stéphane Sanchez, Loubna Echajari, Valérie Friot-Guichard, Philippe Blua, and David Laplanche. Dynamiques des routines organisationnelles en contexte de crise. le cas du département d'information médicale du centre hospitalier de troyes face à la crise Covid-19. *Journal de gestion et d'économie de la santé*, 6(6):364–381, 2021.

[8] Kjeld Schmidt. The problem with 'awareness': Introductory remarks on 'awareness in CSCW'. *Computer Supported Cooperative Work (CSCW)*, 11(3–4):285–298, September 2002.

[9] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green AI. *Communications of the ACM*, 63(12):54–63, December 2020.

[10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] Bastien Sultan, Léon Frénot, Ludovic Apvrille, Philippe Jaillon, and Sophie Coudert. AMULET: A mutation language enabling automatic enrichment of SysML models. *ACM Transactions on Embedded Computing Systems*, 24(3), May 2025. Article 46, 28 pages.

[12] Vivienne Sze, Yu-Hsin Chen, Tien-Ju Yang, and Joel S. Emer. Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 105(12):2295–2329, December 2017.