# Efficient Deployment of AI Applications in the Edge-Network-Cloud Continuum

*Toward Scalable and Sustainable AI Across Heterogeneous Resources*

**Supervisor:** Frédéric Giroire, CNRS Director of Research, 3IA chair holder.
https://www-sop.inria.fr/members/Frederic.Giroire/
frederic.giroire@cnrs.fr

**Laboratory:** COATI team, I3S laboratory (Université Côte d'Azur/CNRS) and Inria (https://team.inria.fr/coati/, https://www.i3s.unice.fr/en/)

**Place:** Centre Inria de l'Université Côte d'Azur, 2004 route des Lucioles, Sophia Antipolis, France (https://www.inria.fr/fr/centre-inria-universite-cote-azur)

## Context and Motivation

The deployment of AI applications is undergoing a paradigm shift with the advent of 5G/6G networks, the Internet of Things (IoT), and edge computing. This evolution enables services to be deployed across the edge-network-cloud continuum [1], leveraging heterogeneous resources from edge devices (e.g., smartphones, microcontrollers) to cloud data centers [2,3,4]. This new paradigm addresses critical challenges such as computing resource constraints, bandwidth limitations, memory availability, and energy efficiency, while introducing new complexities in resource allocation, model deployment, and system optimization.

At the same time, AI models, especially deep neural networks, are becoming increasingly complex, requiring substantial computational power, memory, and energy. For instance, large models often exceed the capabilities of edge devices, while cloud-centric deployments face bandwidth and latency bottlenecks. The need for efficient deployment strategies that balance these constraints is more pressing than ever.

## Scientific Objectives

This thesis aims to develop methods for efficient deployment of AI applications in the edge-network-cloud continuum, addressing resource constraints across computing, bandwidth, memory, and energy. The research will address the following challenges:

### Efficient Deployment Strategies

- **Model Compression:** Investigate techniques such as **quantization, pruning, and knowledge distillation** to reduce the computational and memory footprint of deep learning models without sacrificing accuracy [7, 9, 13].
- **Cascade Systems:** Explore **early-exit architectures** and **multi-stage inference** to dynamically select the most appropriate model (from lightweight to heavyweight)

based on real-time constraints (e.g., battery level, network latency, device memory) [10, 11].
- **Federated Learning:** Study **federated learning (FL)** as a means to distribute AI training and inference across edge devices, reducing the need for data centralization and lowering resource costs (computing, bandwidth, energy) associated with data transfer and cloud compute. FL allows models to be trained locally on devices, with only model updates (not raw data) being communicated, thus improving efficiency and privacy [14].
- **Resource-Aware Scheduling:** Design algorithms to **optimize task placement** (edge vs. cloud) and **scheduling policies** for AI workloads, balancing latency, bandwidth, compute, memory, and energy [13].

### Trade-offs Between Efficiency and Performance

- **Quantitative Analysis:** Measure the resource usage (computing, bandwidth, memory, energy) of AI workloads across different deployment scenarios (edge, network, cloud) and model configurations.
- **Adaptive Configurations:** Develop adjustable models that can be reconfigured on-the-fly to adapt to varying resource constraints and application requirements.

### Operational Impact

- **Resource Footprint Modeling:** Extend existing frameworks to estimate the **resource consumption** (compute, memory, bandwidth, energy) of AI deployments, accounting for both local and distributed execution.
- **Optimization for Scalability:** Propose deployment strategies that minimize resource waste, improve scalability, and ensure reliable performance across heterogeneous environments.

## Research Activities

1. Analyze resource usage of AI deployments in the edge-network-cloud continuum.
2. Design algorithmic methods for efficient scheduling of AI workloads.
3. Investigate trade-offs between resource efficiency and model accuracy in compression techniques.
4. Develop adaptive deployment frameworks using cascade systems and early-exit models.
5. Evaluate environmental impact of proposed methods using lifecycle assessment tools.

## Required Skills and Profile

The ideal candidate should have:

- Knowledge of machine learning, especially neural networks, graph neural networks, or federated learning.
- Strong mathematical, optimization, and algorithmic background.
- Programming expertise in Python, with experience in PyTorch or TensorFlow.
- Familiarity with networking and edge computing (e.g., MEC, IoT, 5G/6G).
- Analytical skills for designing and evaluating optimization algorithms.
- Fluency in English.

# References


[1] H. Hua, Y. Li, T. Wang, N. Dong, W. Li, and J. Cao, “Edge comput- ing with artificial intelligence: A machine learning perspective,” ACM Computing Surveys, vol. 55, no. 9, pp. 1–35, 2023.

[2] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, “When edge meets learning: Adaptive control for resource- constrained distributed machine learning,” in IEEE INFOCOM 2018- IEEE conference on computer communications. IEEE, 2018, pp. 63– 71.

[3] G. Drainakis, P. Pantazopoulos, K. V. Katsaros, V. Sourlas, and A. Amdi- tis, “On the distribution of ml workloads to the network edge and beyond,” in IEEE INFOCOM 2021-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2021, pp. 1–6.

[4] W. Gao, Q. Hu, Z. Ye, P. Sun, X. Wang, Y. Luo, T. Zhang, and Y. Wen, “Deep learning workload scheduling in gpu datacenters: Taxonomy, challenges and vision,” arXiv preprint arXiv:2205.11913, 2022.

[5] J. Lin, W.-M. Chen, J. Cohn, C. Gan, and S. Han, “Mcunet: Tiny deep learning on iot devices,” in Annual Conference on Neural Information Processing Systems (NeurIPS), 2020.

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520.

[7] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, “Model compression and acceleration for deep neural networks: The principles, progress, and challenges,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 126– 136, 2018.

[8] J. Yu and T. Huang, “Autoslim: Towards one-shot architecture search for channel numbers,” 2019. [Online]. Available: https://arxiv.org/abs/1903.11728

[9] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han, “Once-for-all: Train one network and specialize it for efficient deployment,” in International Conference on Learning Representations, 2020. [Online]. Available: https://openreview.net/forum?id=HylxE1HKwS

[10] Viola, P., & Jones, M. (2001, December). Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 (Vol. 1, pp. I-I). Ieee.

[11] Rabanser, S., Rauschmayr, N., Kulshrestha, A., Poklukar, P., Jitkrittum, W., Augenstein, S., ... & Tombari, F. (2025). Gatekeeper: Improving model cascades through confidence tuning. NeurIPS 2025.

[12] Natale, E., Ferré, D., Giambartolomei, G., Giroire, F., & Mallmann-Trenn, F. (2024). On the sparsity of the strong lottery ticket hypothesis. Advances in Neural Information Processing Systems, 37, 40565-40592.

[13] Barros, T. D. S., Giroire, F., Aparicio-Pardo, R., Perennes, S., & Natale, E. (2024, May). Scheduling with fully compressible tasks: Application to deep learning inference with neural network compression. In 2024 IEEE 24th International Symposium on Cluster, Cloud and Internet Computing (CCGrid) (pp. 327-336). IEEE.

[14] Savazzi, S., Rampa, V., Kianoush, S., & Bennis, M. (2022). An energy and carbon footprint analysis of distributed and federated learning. IEEE Transactions on Green Communications and Networking, 7(1), 248-264.

[15] Barros, T. S., Giroire, F., Aparicio-Pardo, R., & Moulierac, J. (2025). Small is Sufficient: Reducing the World AI Energy Consumption Through Model Selection. arXiv preprint arXiv:2510.01889.