

Deploying AI Workloads in Real-Time Systems

A PhD Position

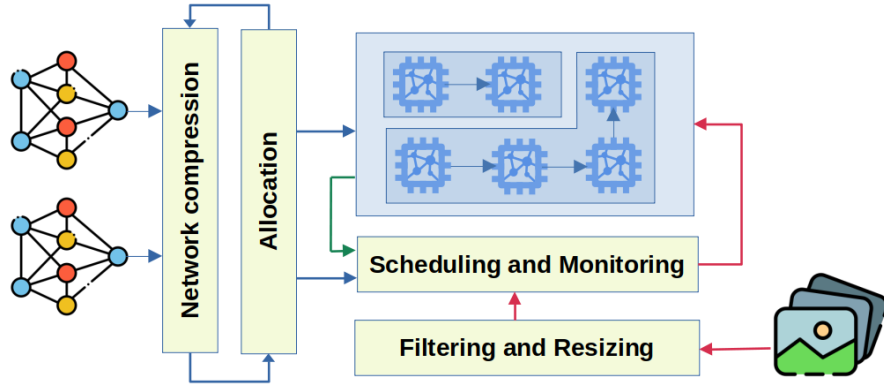
<https://www.laas.fr/fr/offres/deploying-ai-workloads-in-real-time-systems/>

Laboratoire d'analyse et d'architecture des systèmes (LAAS-CNRS), Toulouse, France

Keywords— real-time systems, embedded AI, neural network, hardware accelerators, optimization

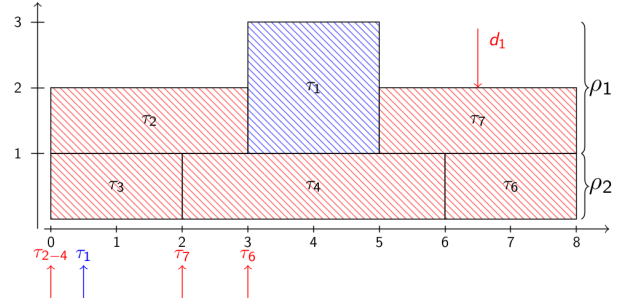
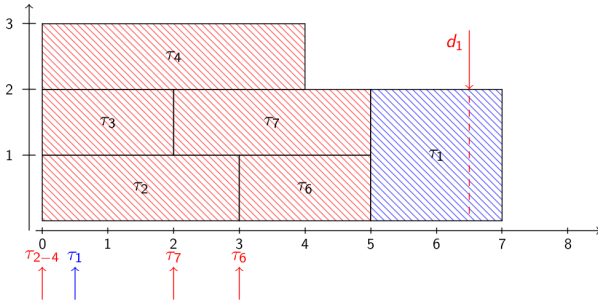
Context. Deploying large AI models on edge devices presents several challenges due to inherent hardware constraints and the rapidly increasing size of AI models. Embedded platforms typically offer limited computational resources and storage capacity, with less flexibility than cloud computing infrastructures. In recent years, numerous hardware solutions specifically designed for edge AI have emerged. These domain-specific architectures exploit parallelism, optimize memory transfers, and support reduced-precision computation. At the software level, specialized techniques aim to reduce model size so that models can fit within the available on-chip memory by reducing the number of parameters and their bit-width representation, through approaches such as model compression, pruning, and quantization.

Real-time scheduling for AI workloads. Nonetheless, hardware acceleration and software optimization alone cannot guarantee real-time performance. Resource management and scheduling are required to decide, at each instant, which neural network should execute on which processing unit. Traditionally, these approaches focus primarily on meeting timing objectives, whereas edge AI requires consideration of several additional aspects: accuracy, energy consumption, and, when necessary, thermal constraints. Satisfying these constraints requires accurately characterizing the AI workload (i.e., the task model) in terms of its parallelism level, memory footprint, and precision. Consequently, scheduling policies should not only account for these additional constraints but also be designed with regard to the underlying hardware capabilities (e.g., large number of processing units, limited on-chip memory capacity, and high context-switching overhead).



Objectives. Real-time systems are composed of multiple tasks, each with specific timing constraints (e.g., deadlines and execution frequency), execution requirements, and criticality levels. In edge AI applications, each real-time task can run a neural network inference. The goal of this thesis is to propose a framework for integrating AI models into real-time and edge AI systems while meeting specific timing and performance constraints on modern AI hardware accelerators. To this end, the following objectives will be addressed:

- **Design space exploration:** Develop a toolbox to optimize the compression levels of neural networks under timing and performance constraints.
- **Hardware resource partitioning and configuration:** Determine the optimal partitioning of available hardware resources, find the best topologies of processing elements (e.g., pipeline depth), and assign scheduling parameters (e.g., priorities and preemption points).
- **Real-time scheduling:** Develop scheduling policies tailored to AI models and AI hardware accelerators.
- **Runtime-adaptive neural networks:** Develop neural network models with runtime control of their execution, enabling the workload to adapt to current system utilization, available parallelism, or energy constraints.



Research group. The project is a collaboration among the *Laboratory for Analysis and Architecture of Systems (LAAS-CNRS)*, the *Laboratory for Computer Science in Images and Information Systems (LIRIS-CNRS)*, *Inria Center at Rennes University*, and the *Technical University of Munich (TUM)*. The host laboratory is the LAAS-CNRS in Toulouse, France. The thesis will be co-supervised by Dr. Tomasz Kloda (*LAAS-CNRS*), Prof. Stefan Duffner (*LIRIS-CNRS*), Prof. Angeliki Kritikakou (*Inria Rennes*) and Dr. Binqi Sun (*TUM*).

Contact. For inquiries about the position, please contact Dr. Tomasz Kloda (tomasz.kloda@laas.fr).

Qualifications. Candidates must hold a graduate degree (or equivalent) in *Computer Science* or *Mathematics*.

Duration. 36 months. The position is available immediately.

General Literature

- [1] John A. Stankovic. “Misconceptions about Real-Time Computing: A Serious Problem for Next-Generation Systems”. In: *Computer* 21.10 (Oct. 1988), pp. 10–19. DOI: [10.1109/2.7053](https://doi.org/10.1109/2.7053).
- [2] Yizeng Han et al. “Dynamic Neural Networks: A Survey”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2022), pp. 7436–7456. DOI: [10.1109/TPAMI.2021.3117837](https://doi.org/10.1109/TPAMI.2021.3117837).

Our Recent Publications

- [3] B. Sun et al. “Partitioned Scheduling and Parallelism Assignment for Real-Time DNN Inference Tasks on Multi-TPU”. In: *Proceedings of the 61st ACM/IEEE Design Automation Conference (DAC)*. 2024. DOI: [10.1145/3649329.3655979](https://doi.org/10.1145/3649329.3655979). URL: <https://dl.acm.org/doi/10.1145/3649329.3655979>.
- [4] B. Sun, T. Kloda, and M. Caccamo. “Strict Partitioning for Sporadic Rigid Gang Tasks”. In: *Proceedings of the IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. 2024. DOI: [10.1109/RTAS61025.2024.00028](https://doi.org/10.1109/RTAS61025.2024.00028).
- [5] B. Sun et al. “SAPar: A Surrogate-Assisted DNN Partitioner for Efficient Inferences on Edge TPU Pipelines”. In: *Proceedings of the International Conference on Embedded Software (EMSOFT)*. 2025. DOI: [10.1145/3761813](https://doi.org/10.1145/3761813). URL: <https://dl.acm.org/doi/10.1145/3761813>.
- [6] B. Sun et al. “AI Inference in the Heat: Thermal-Aware Strict Partitioning for Configurable Real-Time Gang Tasks”. In: *Proceedings of the IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. 2026. DOI: [10.1109/RTAS68450.2026.00031](https://doi.org/10.1109/RTAS68450.2026.00031).
- [7] M. Zouhdi et al. “Pruning of Deep Neural Networks for Real-Time Execution on Edge TPU”. In: *Proceedings of the Euromicro Conference on Digital System Design (DSD)*. To appear. 2026.