



HAL
open science

Study and design of MRAM-based in-memory computing architectures for machine learning applications

Bruno Lovison-Franco, Aymen Romdhane, David Novo, Pascal Benoit, Guillaume Prenat, Lorena Anghel

► **To cite this version:**

Bruno Lovison-Franco, Aymen Romdhane, David Novo, Pascal Benoit, Guillaume Prenat, et al.. Study and design of MRAM-based in-memory computing architectures for machine learning applications. Journées Scientifiques Nationales 2024 du PEPR Électronique, Mar 2024, Grenoble, France. lirmm-04717768

HAL Id: lirmm-04717768

<https://hal-lirmm.ccsd.cnrs.fr/lirmm-04717768v1>

Submitted on 2 Oct 2024

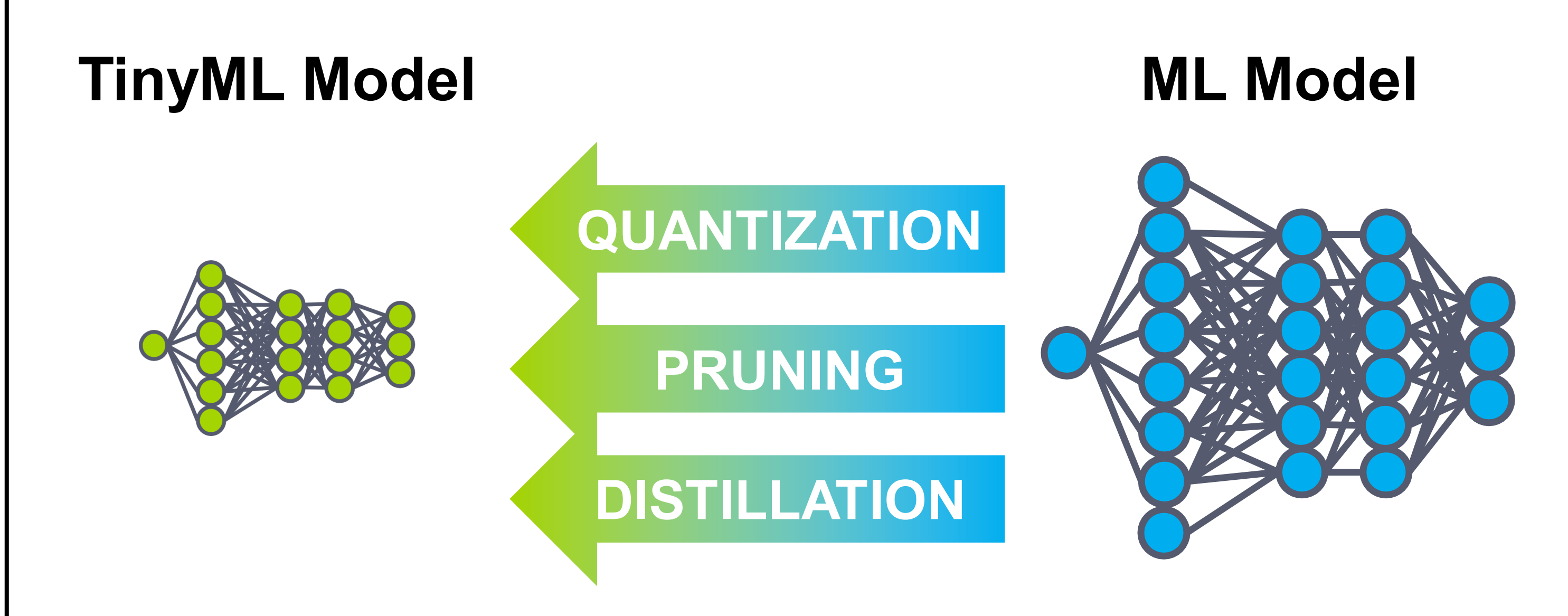
HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

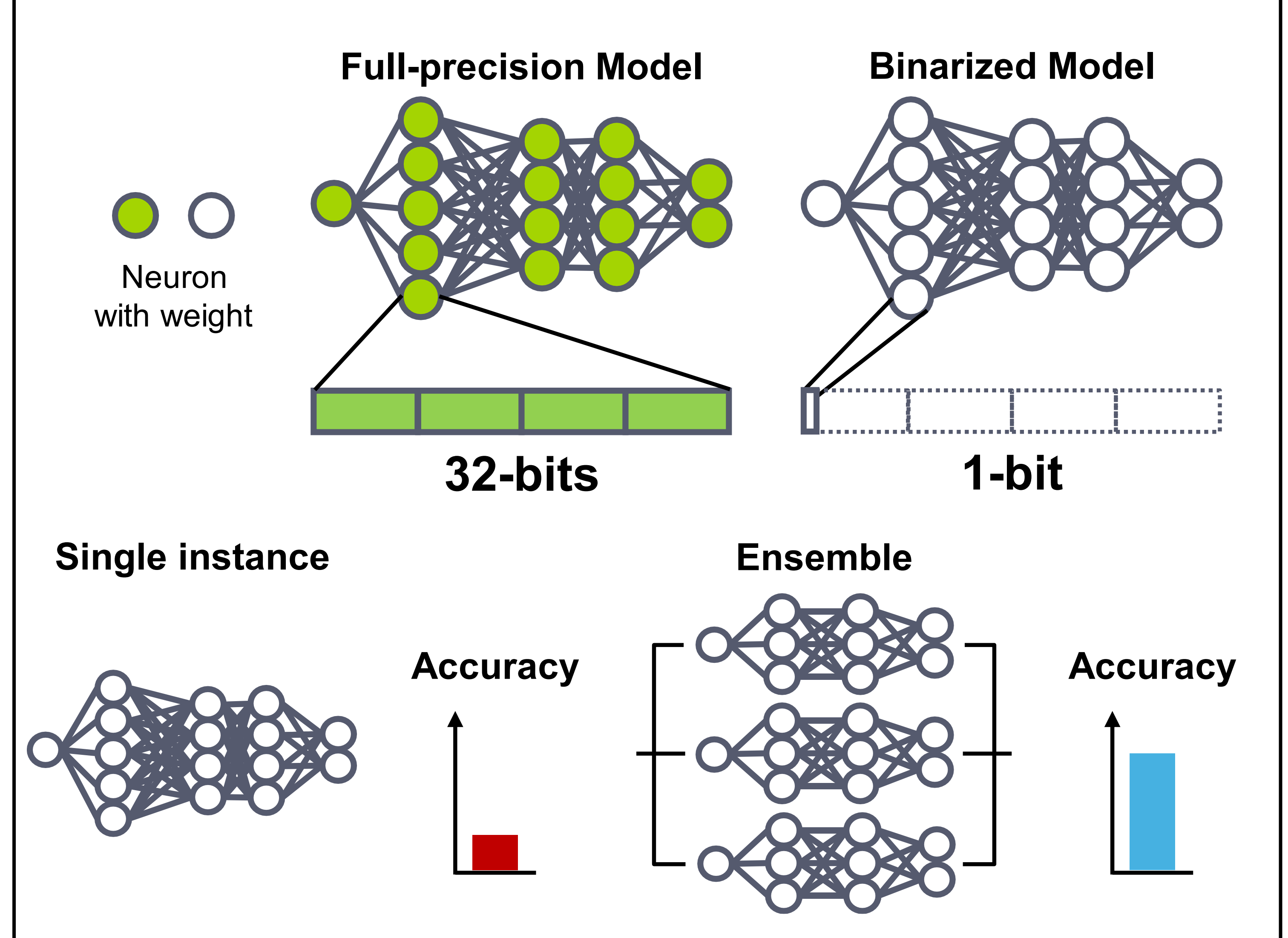
Study and design of MRAM-based in-memory computing architectures for machine learning applications

Bruno Lovison Franco¹, Aymen Romdhane², David Novo¹, Pascal Benoit¹, Guillaume Prenat², Lorena Anghel²

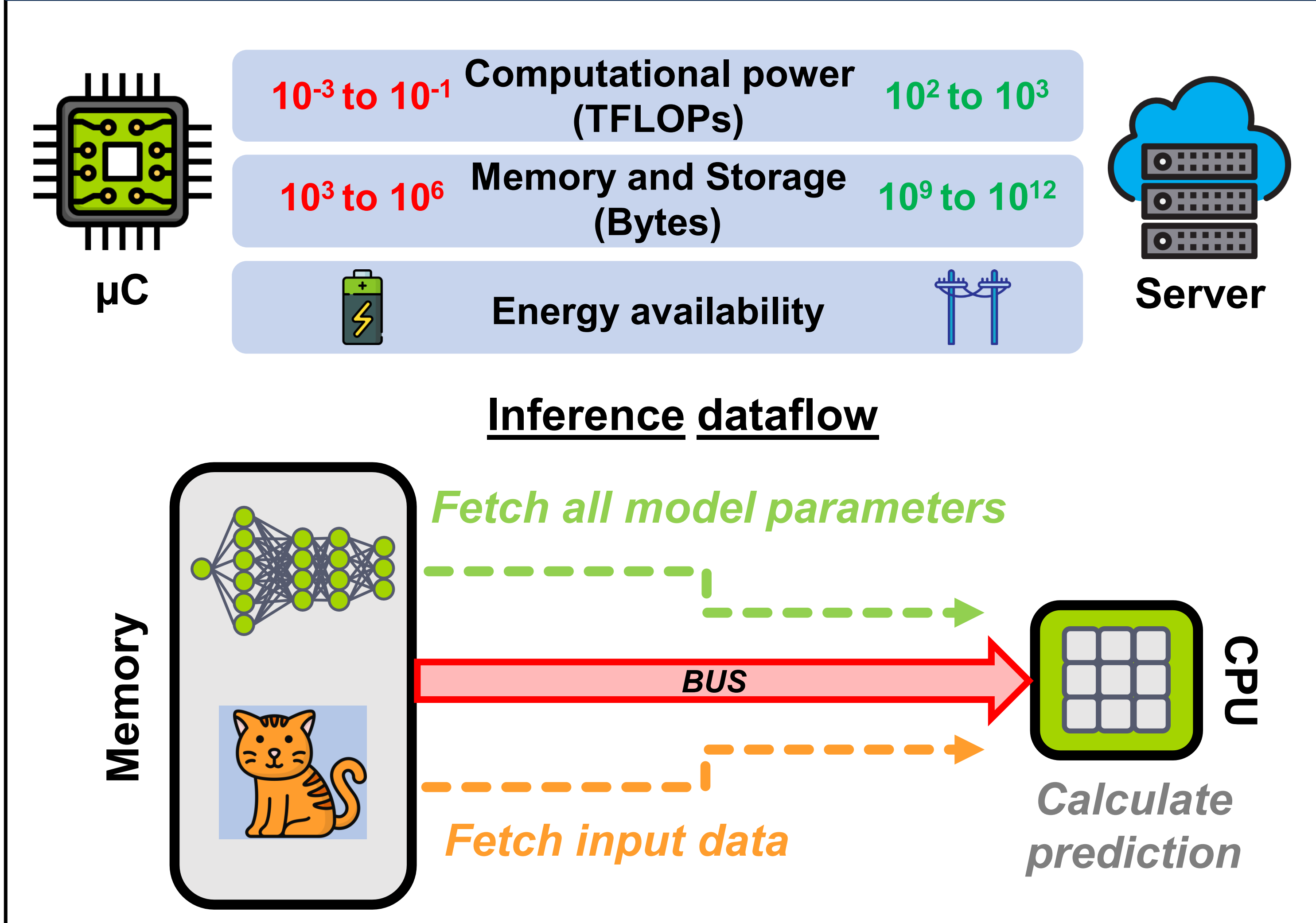
① from regular ML models towards TinyML



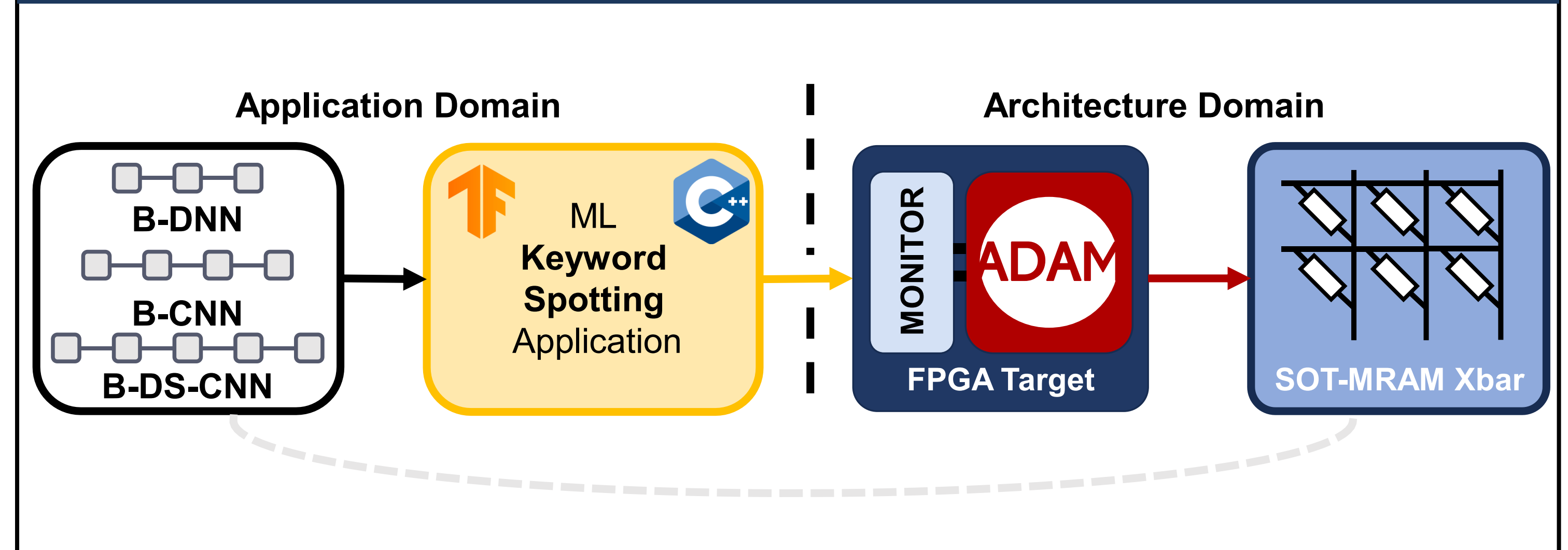
③ Binary Neural Networks & Ensembles



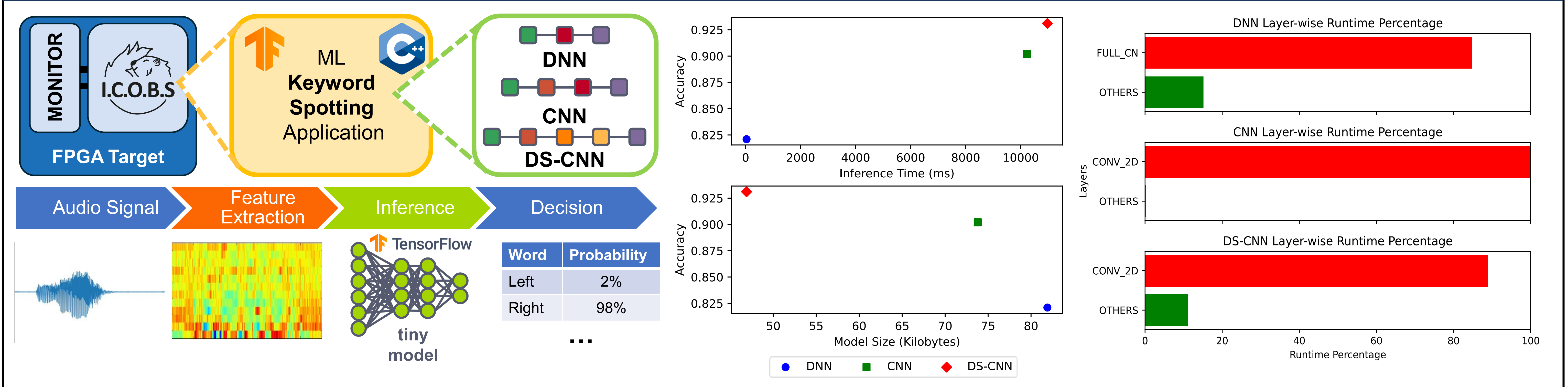
② Model Size, Data Movement and Operation Latency



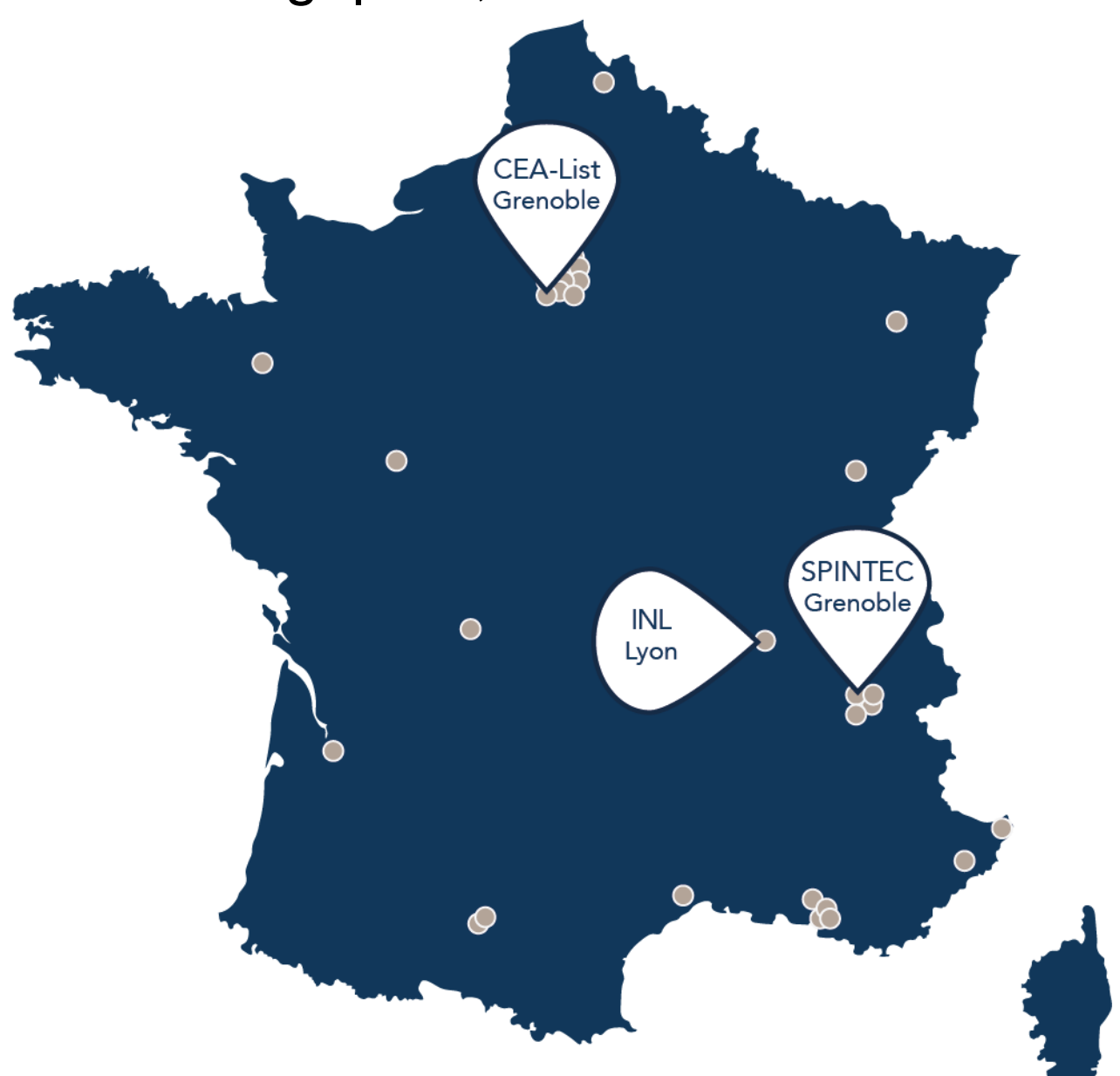
④ Network and Architecture Co-optimization



⑤ Keyword Spotting Application and Bottlenecks



1. Tchendjou, Ghislain Takam, et al. "Spintronic Memristor based Binarized Ensemble Convolutional Neural Network Architectures." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (2022).
2. Zhu, Shilin, Xin Dong, and Hao Su. "Binary ensemble neural network: More bits per network or more networks per bit?." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
3. Mutlu, Onur, et al. "A modern primer on processing in memory." Emerging Computing: From Devices to Systems: Looking Beyond Moore and Von Neumann. Singapore: Springer Nature Singapore, 2022. 171-243.



- (1) Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, Université de Montpellier, CNRS -161, rue Ada, 34095 Montpellier cedex 5 - firstname.lastname@lirmm.fr
- (2) Univ. Grenoble Alpes, CEA, CNRS, Grenoble INP, SPINTEC, 38000 Grenoble, France