Flavio Ponzina

Assistant Professor
Department of Computer Science and Engineering
San Diego State University

Email: fponzina@sdsu.edu
Languages: Italian (native), English (fluent), French (B2), Spanish (B1)
Scholar: Flavio Ponzina - Google Scholar
LinkedIn: https://www.linkedin.com/in/flavio-ponzina-7b51a7139

INTERESTS AND EXPERTISE Edge A.I., Embedded systems, HW-SW co-design optimization, Memory-constrained ensemble learning, Processing In-Memory

ACADEMIC BACKGROUND

Aug. 2025 – **San Diego State University (SDSU), United States of America**

- Assistant professor - Department of ECE

Oct. 2023 – Aug 2025. **University of California San Diego (UCSD), United States of America**

- Postdoctoral scholar at UC San Diego. Supervisor: Prof. Tajana Rosing

Aug. 2019 – Sep 2023 **École Polytechnique Fédérale de Lausanne (EPFL), Switzerland**

- Ph.D. in Electrical and Electronic Engineering. Supervisor: Prof. David Atienza
  Thesis: *Hardware-Software Co-Design Methodologies for Edge AI Optimization*

Sep. 2016 – Dec. 2018 **Politecnico di Torino, Italy**

- MS in Computer Engineering, 110/110 *cum laude*
  Thesis: *Hardware-Aware Optimization of Embedded Convolutional Neural Networks*

CONFERENCE PAPERS

- Chien-Yi Yang, Sheng Chen, Minxuan Zhou, **Flavio Ponzina**, Dongxia Wu, Raid Ayoub, Pietro Mercati, Mahesh Subedar, Yian Ma, Rose Yu, Tajana Rosing, "*Multi-Objective Bayesian Optimization for Efficient HDnn-PIM Software-Hardware Co-Design with Metric Constraints*", under review
- You Hak Lee, Xiaofan Yu, Quanling Zhao, **Flavio Ponzina**, and Tajana Rosing, "*FedUHD: Unsupervised Federated Learning with Hyperdimensional Computing*", under review
- Sumukh Pinge, Ashkan Moradifirouzabadi, Keming Fan, Prasanna Venkatesan Ravindran, Tanvir H. Pantha, Zheyu Li, Weihong Xu, Zihan Xia, **Flavio Ponzina** et. al "*FeNOMS: Enhancing Open Modification Spectral Library Search with In-Storage Processing of Ferroelectric NAND (FeNAND) Flash*", under review
- Shriniwas Kulkarni, **Flavio Ponzina**, Tajana Rosing, *"HyperDrone - an Accurate, Robust, Fast, and Energy-Efficient Approach for Drone Classification",* under review
- *Siqi Li*, **Flavio Ponzina**, *Tajana Rosing, "AA-DiT An Algorithm-Architecture Co-Design for Diffusion Tranformer Acceleration",* under review
- Tianqi Zhang, **Flavio Ponzina**, Tajana Rosing, *"TRQ - Tiered Residual Quantization for LLM Vector Search in Far-Memory-Aware ANN Systems",* under review
- Jangseon Park, **Flavio Ponzina**, Tajana Rosing, *"Energy-Efficient LLM Inference with Hybrid SRAM and MRAM on Edge Device",* under review
- Nilesh Prasad Pandey, Shriniwas Kulkarni, David Wang, Onat Gungor, **Flavio Ponzina,** and Tajana Rosing, "*DPQ-HD: Post-Training Compression for Ultra-Low Power Hyperdimensional Computing*", GLSVLSI, 2025
- **Flavio Ponzina**, Sumukh Pinge, Zheyu Li, Abhijay Deevi, Yilin Ge, Mingu Kang, and Tajana Rosing *"SmartMS - Efficient Hierarchical Database Search for Mass Spectrometry via Processing-in-Memory",* ISLPED, 2025
- Le Zhang, Quanling Zhao, Run Wang, Shirley Bian, Onat Gungor, **Flavio Ponzina,** and Tajana Rosing, "*Offload Rethinking by Cloud Assistance for Efficient Environmental Sound Recognition on LPWANs*", SenSys, 2025
- Le Zhang, Onat Gungor, **Flavio Ponzina**, and Tajana Rosing., "*E-QUARTIC: Energy Efficient Edge Ensemble of Convolutional Neural Networks for Resource-Optimized Learning",* ASPDAC, 2025
- Matilda Gaddi, **Flavio Ponzina,** Fatemeh Asgarinejad, Barish Aksanli, and Tajana Rosing. *"HyperECG: ECG Signal Inference from Radar with Hyperdimensional Computing*", BIBE, 2024

- Kumar Ashwani, Yucheng Zhou, Sai Praneeth Potladurthy, Jeoghoon Kim, Weihong Xu, **Flavio Ponzina**, Seounghyun Kim, Ertugrul Cubukcu, Tajana Rosing, Gert Cauwenbergh, and Duygu Kuzum. *"Filament-free Bulk RRAM with High Endurance and Long Retention for Neuromorphic Few-Shot Learning On-Chip*", IEDM, 2024
- **Flavio Ponzina**, Rubén Rodríguez Álvarez, José Miranda Calero, Mathieu Salzmann, Jacques Viertl, Tajana Rosing, Miguel Peón-Quirós, and David Atienza, "*Using ensemble learning to improve radiation tolerance of CNNs in Space Applications*", SPAICE, 2024
- Asgarinejad Fatemeh, **Flavio Ponzina**, Onat Gungor, Tajana Rosing, and Barish Aksanli. "*HDXpose Harnessing Hyperdimensional Computing's Explainability for Adversarial Attacks",* ICCAD, 2024
- Chien-Yi Yang, Minxuan Zhou, **Flavio Ponzina**, Suraj Sathya Prakash, Raid Ayoub, Pietro Mercati, Mahesh Subedar, and Tajana Rosing*, "Multi-Objective Software-Hardware Co-Optimization for HD-PIM via Noise-Aware Bayesian Optimization",* ICCAD, 2024
- **Flavio Ponzina**, Rishikanth Chandrasekaran, Anya Wang, Seiji Minowada, Siddharth Sharma, and Tajana Rosing*. "Multi-Model Inference Composition of Hyperdimensional Computing Ensembles",* ICCD, 2024
- Jason Kong, Lanxiang Hu, **Flavio Ponzina**, and Tajana Rosing. "*TinyAgent: Quantization-aware Model Compression and Adaptation for On-device LLM Agent Deployment.*" In Workshop on Efficient Systems for Foundation Models II@ ICML, 2024.
- Amirhossein Shahbazinia, **Flavio Ponzina**, José Angel Miranda Calero, Jonathan Dan, Giovanni Ansaloni, and David Atienza Alonso. "*Resource-Efficient Continual Learning for Personalized Online Seizure Detection*." In 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2024.
- **Flavio Ponzina**, and Tajana Rosing. "*MicroHD: An Accuracy-Driven Optimization of Hyperdimensional Computing Algorithms for TinyML systems.*" tinyML, 2024.
- Pengbo Yu, **Flavio Ponzina**, Alexandre Sébastien Julien Levisse, Biswas Dwaipayan, Ansaloni Giovanni, Atienza David, and Catthoor Francky. "*DBFS: Dynamic Bitwidth-Frequency Scaling for Efficient Software-defined SIMD*", ISVLSI, 2024.
- Marco Rios**, Flavio Ponzina**, Giovanni Ansaloni, Alexandre Levisse, and David Atienza. "*Error resilient in-memory computing architecture for cnn inference on the edge*." GLSVLSI, 2022.
- **Flavio Ponzina**, Miguel Peon-Quiros, Giovanni Ansaloni, and David Atienza. "*An accuracy-driven compression methodology to derive efficient codebook-based cnns*.", COINS, 2022
- Marco Rios, **Flavio Ponzina**, Giovanni Ansaloni, Alexandre Levisse, and David Atienza. "*Running efficiently cnns on the edge thanks to hybrid sram-rram in-memory computing*.", DATE, 2021
- **Flavio Ponzina**, Marco Rios, Giovanni Ansaloni, Alexandre Levisse, and David Atienza. "*A flexible in-memory computing architecture for heterogeneously quantized CNNs*.", ISVLSI, 2021

JOURNAL PAPERS

- Corey Lammie, **Flavio Ponzina**, Yuxuan Wang, Joshua Klein, Marina Zapater, Irem Boybat, Abu Sebastian, Giovanni Ansaloni, and David Atienza. "*LionHeart: A Layer-based Mapping Framework for Heterogeneous Systems with Analog In-Memory Computing Tiles.*" IEEE TETC, 2025.
- Keming Fan, Ashkan Moradifirouzabadi, Xiangjin Wu, Zheyu Li, **Flavio Ponzina**, Anton Persson, Vikram Adve, Eric Pop, Tajana Rosing, and Mingu Kang*. "SpecPCM: A Low-power PCM-based In-Memory Computing Accelerator for Full-stack Mass Spectrometry Analysis." IEEE JXCDC, 2024.*
- **Flavio Ponzina**, Mialyssa Gomez, Congge Xu, and Tajana Rosing. *"GlucoseHD Predicting Glucose Levels using Hyperdimensional Computing",* IEEE Design and Test, 2024.
- Pengbo Yu**, Flavio Ponzina,** Alexandre Levisse, Mohit Gupta, Dwaipayan Biswas, Giovanni Ansaloni, David Atienza, and Francky Catthoor. "An Energy Efficient Soft SIMD Microarchitecture and Its Application on Quantized CNNs." IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2024.
- Marco Rios, **Flavio Ponzina**, Alexandre Levisse, Giovanni Ansaloni, and David Atienza. "Bit-line computing for CNN accelerators co-design in edge AI inference." IEEE Transactions on Emerging Topics in Computing 11, no. 2, 2023.
- Silvio Zanoli, **Flavio Ponzina**, Tomás Teijeiro, Alexandre Levisse, and David Atienza. "An Error-Based Approximation Sensing Circuit for Event-Triggered Low-Power Wearable Sensors." IEEE Journal on Emerging and Selected Topics in Circuits and Systems 13, no. 2, 2023.
- **Flavio Ponzina**, Marco Rios, Alexandre Levisse, Giovanni Ansaloni, and David Atienza. "*Overflow-free compute memories for edge AI acceleration.*" ACM Transactions on Embedded Computing Systems, 2023.

- **Flavio Ponzina**, Giovanni Ansaloni, Miguel Peón-Quirós, and David Atienza. "*Using algorithmic transformations and sensitivity analysis to unleash approximations in CNNs at the edge*." Micromachines 13, no. 7, 2022.
- **Flavio Ponzina**, Simone Machetti, Marco Rios, Benoît Walter Denkinger, Alexandre Levisse, Giovanni Ansaloni, Miguel Peón-Quirós, and David Atienza. "*A hardware/software co-design vision for deep learning at the edge.*" IEEE Micro 42, no. 6, 2022.
- **Flavio Ponzina**, Miguel Peon-Quiros, Andreas Burg, and David Atienza. "*E2cnns: Ensembles of convolutional neural networks to improve robustness against memory errors in edge-computing devices*." IEEE Transactions on Computers 70, no. 8, 2021.
- Benoît W. Denkinger, **Flavio Ponzina**, Soumya S. Basu, Andrea Bonetti, Szabolcs Balási, Martino Ruggiero, Miguel Peón-Quirós, Davide Rossi, Andreas Burg, and David Atienza. "*Impact of memory voltage scaling on accuracy and resilience of deep learning based edge devices.*" IEEE Design & Test 37, no. 2, 2019.

HONORS & AWARDS

- SRC JUMP 2.0 PRISM annual review, *"Theme 1 - Systems & Software"* Best poster award, 2024
- SRC JUMP 2.0 Undergrad Symposium 2024, Best Mentor Award, 2024
- EPFL School of Engineering Teaching Assistant Award nominee, 2023
- FETCH23, Best presentation award at "*My thesis in 180 seconds*" contest, 2023
- EPFL EDEE Thesis Distinction Candidate, 2023
- ACM ESWEEK 2023. Best Paper Award Nominee, 2023

RESEARCH TALKS

- "*Hardware-Software co-design for ultra-low power edge AI*", Department of Computer Science, Politecnico di Torino, Italy. Nov. 2024.
- "*Ultra-Efficient Edge AI with Energy-Aware Ensembles and Hyperdimensional Computing*", SRC JUMP 2.0 CoCoSys Theme Meeting – Hardware/Software co-design, Sep. 2024.
- "*Ultra-Efficient Edge AI with Energy-Aware Ensembles and Hyperdimensional Computing*", Department of Computer Science, Cambridge University, UK. Sep. 2024.

INDUSTRY COLLABORATIONS

- Intel. "*Frontend-Backend co-optimization for HD-PIM architectures*", 2024
- Samsung. "*Leveraging MRAM for energy-efficient search and storage of vector databases*", 2024
- IBM. "*Hybrid digital/analog PIM mapping of CNN layers for efficient inference"*, 2023
- ClearSpace/ESA. "*Improving radiation tolerance of CNN-based spacecraft navigation system"*, 2022

ACADEMIC SERVICE

- TPC member of the *LatinIoT-2025* conference, Fortaleza, Brazil, 2025.
- Session Chair at ICCD'24, "*Session 10B: Accelerators*", 2024
- Journal paper reviewer for ISCAS, DATE, ASPLOS, and IEEE Transactions on Computers, TCAD, and ACM CSUR
- EDEE Student Committee representative, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland (2020-2023)

FUNDING

- DARPA & SRC JUMP 2.0 funding for PRISM projects, $300K, 2025
- DARPA & SRC JUMP 2.0 funding for PRISM projects, $400K, 2024

TEACHING EXPERIENCE

San Diego State University (SDSU)

- COMPE 361. Advanced Programming
- COMPE 271. Computer Organization

University of California San Diego (UCSD)

- CSE 147. Introduction to Embedded Systems
- CSE 237a. Introduction to Embedded Computing

École Polytechnique Fédérale de Lausanne (EPFL)

- Microprogrammed Embedded Systems - Teaching Assistant