UNIVERSITÀ DEGLI STUDI DI NAPOLI
FEDERICO II

itee PhD
information technology
electrical engineering

Vincenzo Maisto

Harnessing Energy cross the Computing Landscape:
Cross-domain Approaches for Energy efficiency and Hardware Consolidation

Tutor: Alessandro Cilardo

Cycle: XXXVII Year: 2022

itee PhD
information technology
electrical engineering

Candidate's information

- MSc degree: Computer Engineering at UNINA
- Research group/laboratory: SECLab / HPC
- PhD start date: 1st January 2022 – 31st December 2024
- Scholarship type: MUR PON
- Partner company under DM 1061: A3cube Inc.
- Periods abroad: 6 months visit at ETH Zurich

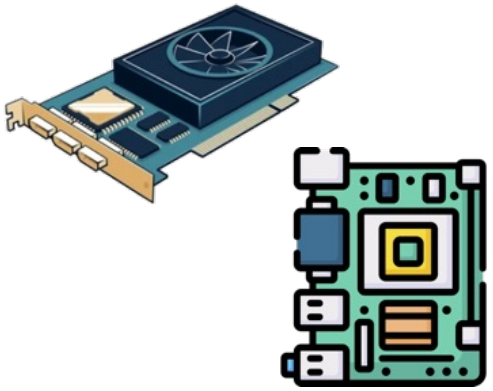
Summary of study activities

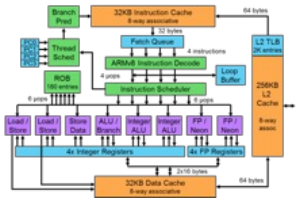
- Ad hoc PhD courses / schools:
  - Virtualization Technologies and their applications, ITEE, UNINA
  - Big Data Analytics and Architectures, ITEE, UNINA
  - Academic Entrepreneurship, DII, UNINA
  - Statistical data analysis for science and engineering research, ITEE, UNINA
  - Strategic Orientation for Stem Research & Writing, ITEE, UNINA
- Courses borrowed from MSc curricula:
  - Introduction to Quantum Circuits, DIETI, UNINA
  - Quantum Information, DIETI, UNINA
- Conferences / events attended:
  - DATE 2022
  - National Workshop for Technology Transfer and Higher Education 2022
  - QUATIC 2022
  - “10 Years of PULP” workshop
  - ITADATA 2024

Research topic

Computing Architectures

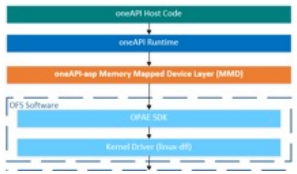
- **Heterogeneity in MPSoCs:**
  - Server-class acceleration cards
  - Edge MPSoC platforms
  - SoC and CPU design

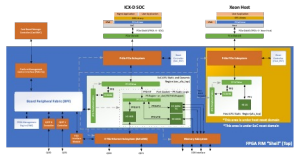




- **Green Computing**
  - Energy efficiency and low-power design
  - Scalability and Heterogeneity

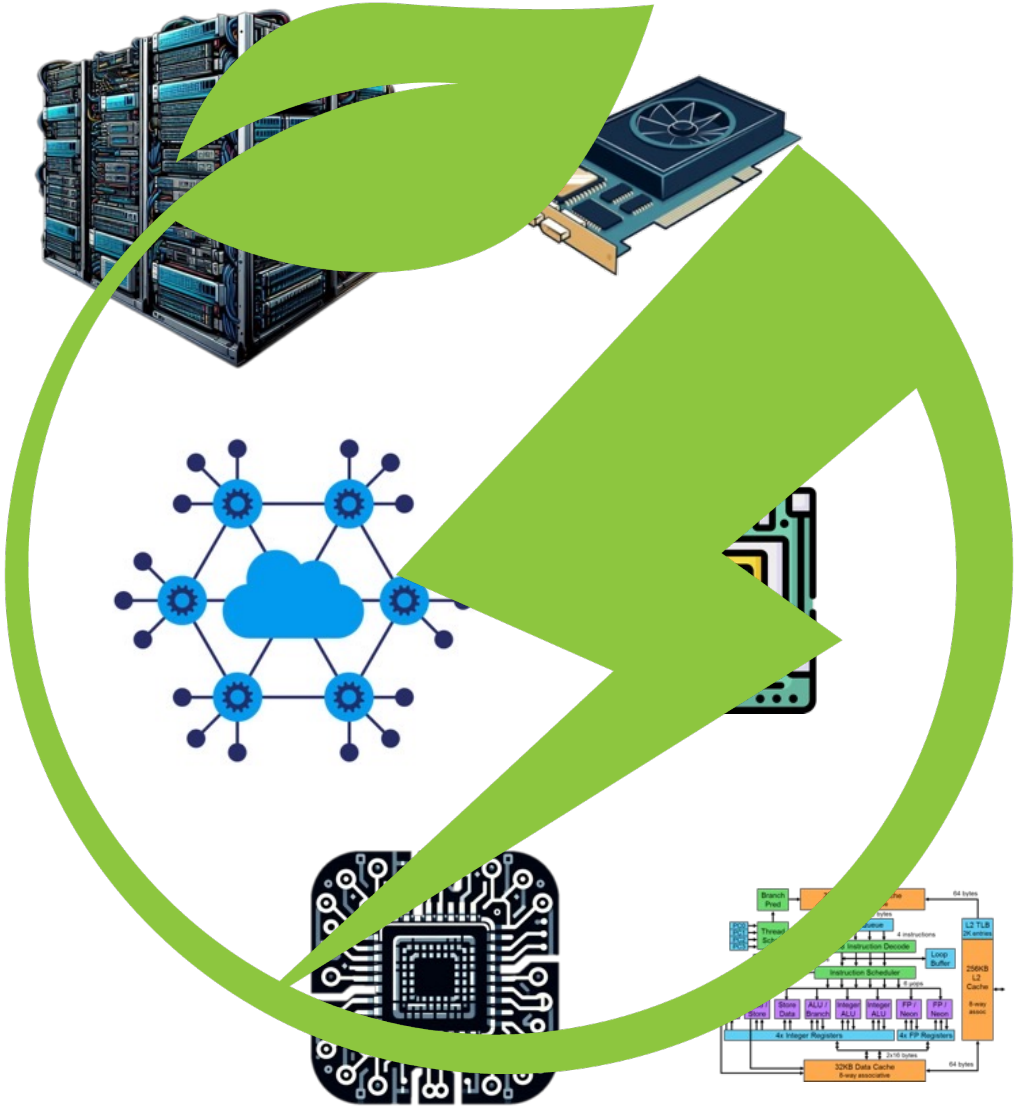
- **Advanced Hardware/Software Co-design**





The Digital Computing Landscape

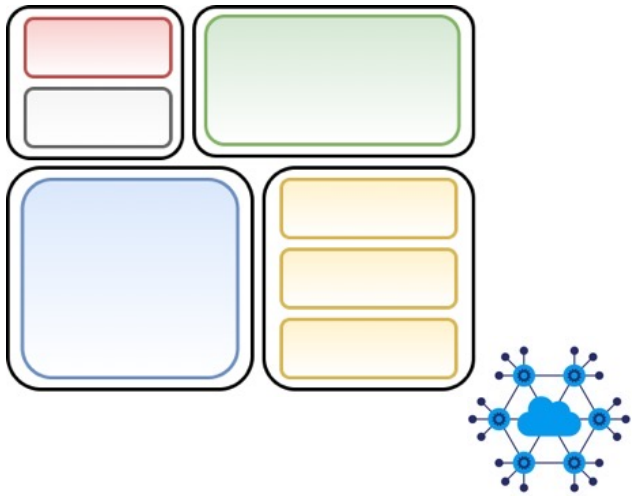
- HPC / Cloud computing
  - Server-class
  - Cluster-based
  - **Business-oriented**
- Edge Computing
  - Field-deployment
  - Driver for IoT
  - **Battery-limited**
- SoC and CPU design
  - Micro-architectural design
  - System architecture
  - **Power-limited**

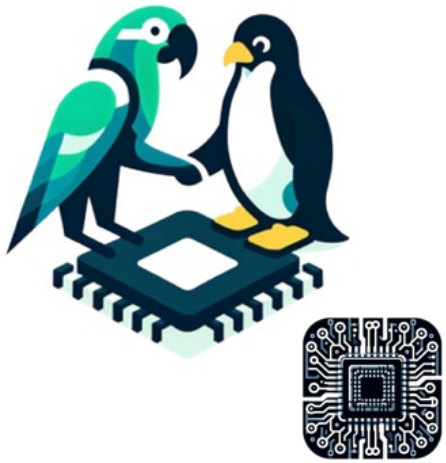


Research results

In the quest for **energy efficiency** across computing domains, this thesis presents:

1. **Common cross-domain** computer design approaches
2. Enabling **hardware consolidation** and **multi-threading/tenancy**
3. **Cross-domain system design and evaluation**





Research Products

[P1]	V. Maisto and A. Cilardo, A Pluggable Vector Unit for RISC-V Vector Extension, **DATE 2022,** Antwerp, Belgium, 2022, pp. 1143-1148, doi: 10.23919/DATE54114.2022.9774501
[P2]	Cilardo, A., Maisto, V., Mazzocca, N., Rocco di Torrepadula, F. (2022). A Proposal for FPGA-Accelerated Deep Learning Ensembles in MPSoC Platforms Applied to Malware Detection. **QUATIC 2022** doi: https://doi.org/10.1007/978-3-031-14179-9_16
[P3]	Cilardo, A., Maisto, V., Mazzocca, N., & Rocco di Torrepadula, F. (2023). An approach to the systematic characterization of multitask accelerated CNN inference in edge MPSoCs. **ACM Transactions on Embedded Computing Systems.** DOI: https://doi.org/10.1145/3611015
[P4]	Vincenzo Maisto, Matteo Perrotti, Moritz Imfeld GitHub PRs for open-source PULP products: • Ara integration in Cheshire (1) • Ara integration in Cheshire (2) • [HW/SW] Cheshire integration - Linux on FPGA

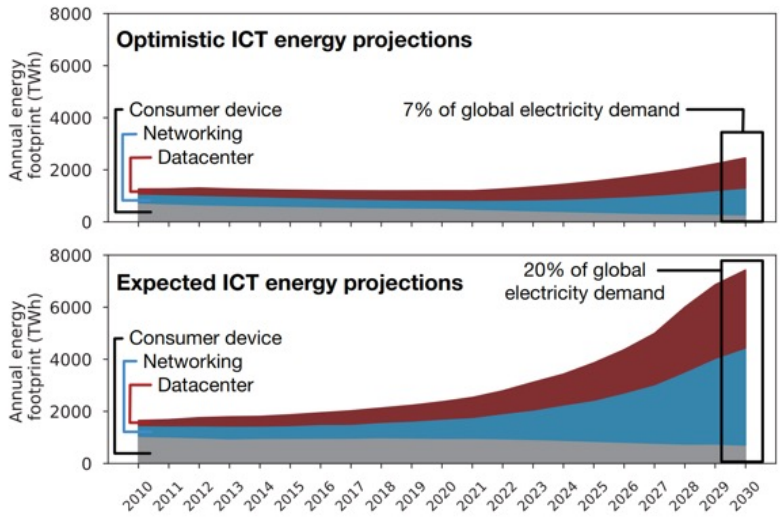
Research Products

[P5]	Rocco di Torrepadula, F., Maisto, V., Cilardo, A., Mazzocca, N. Distilling Knowledge for Low-Carbon AIoT [under review at **IEEE TSUSC**]
[P6]	Maisto, V., Cilardo, A., Billi, E., Fader, C. Extended Abstract: Multi-threaded FPGA Offloading of Erasure Coding for Hadoop Distributed File System [Presented at **BigHPC ITADATA 2024**]
[P7]	V. Maisto and A. Cilardo Multi-DPU: an Energy efficient Hardware/Software Architecture for Multi-tenant Deep Learning Acceleration on Edge MPSoCs [submitted at **DATE 2024**]
[P8]	Maisto, V., Cilardo, A., Billi, E., Fader, C. A Hardware/Software Architecture for Multi-threaded Offloading of Erasure Codes in Distributed File Systems. [submitted at **HPCA 2025**]

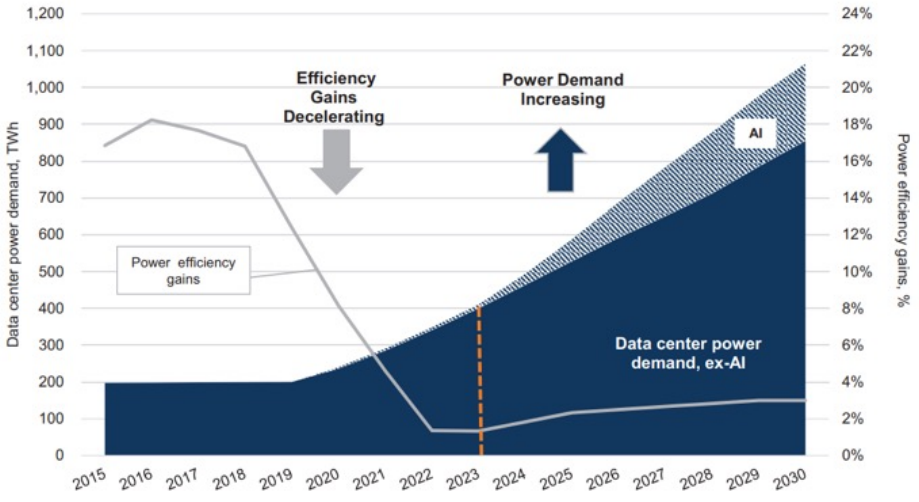
The Environmental Footprint of Computing

Computing technologies are nowadays ubiquitous and expanding

- Computing alone could account for **20%** of the global energy demand by 2030
- "*AI is poised to drive 160% increase in data center power demand*", Goldman Sachs, 2024
- Interest in **Green Computing** approaches is rising, e.g.:
  - **Green AI** Schwartz et al. 2020. Green AI. Commun. ACM (December 2020). https://doi.org/10.1145/3381831
  - **Water consumption awareness** Li et al. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models, 2024, https://doi.org/10.48550/arXiv.2304.03271
  - …



U. Gupta *et al.*, "Chasing Carbon: The Elusive Environmental Footprint of Computing," *HPCA 2021* https://doi.org/10.1109/HPCA51647.2021.00076

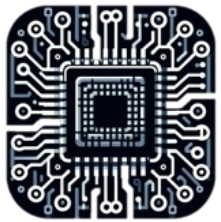


Masanet et al. (2020), IEA, Cisco, Goldman Sachs Global Investment Research

Harnessing Energy Across the Computing Landscape

Modern computing technologies are extremely **diverse and heterogeneous**

- Most techniques are **domain-specific** and non-generalizable
- Common approaches are challenging to find and even more difficult to deploy
- The landscape is diverse in **scale, nature and requirements**



There is no silver bullet to save energy → **We need to craft it!**

This thesis enables and validates
common approaches for energy efficient computing
across computing domains

Cross-domain Approaches for Energy Efficiency and Heterogeneous Multi-threading

Searching for a **common approach**,

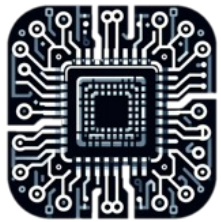
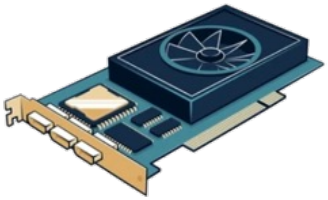
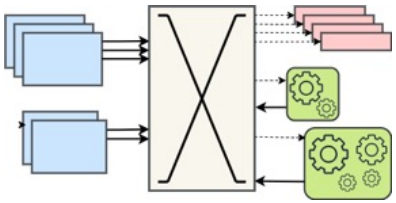
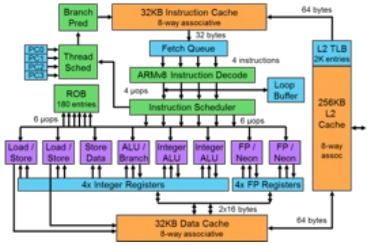
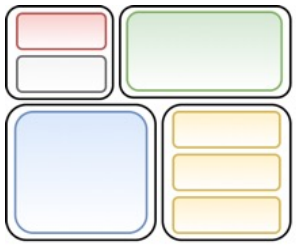
this thesis avoids **domain-specific methodologies**, e.g.:

Low-power electronic design	Battery technologies
Voltage and frequency scaling	Rack-level cooling optimizations

As common strategies, we delve into:

Hardware/software acceleration	→	*Save time, save energy*
Heterogeneity	→	*Right tools for the job*
Hardware consolidation	→	*Max' compute per device*
Energy efficiency	→	*Max' compute per Joule*

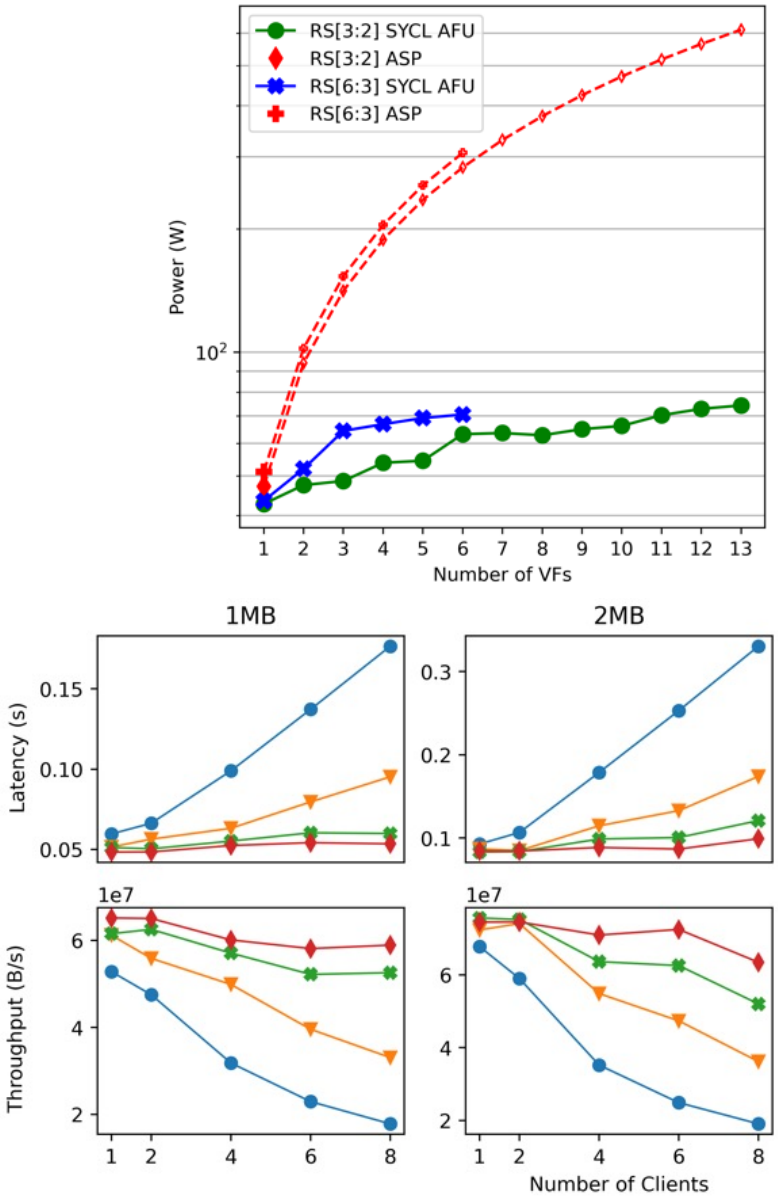
Our Target Computing Landscape

Domain	**Cloud** 	**Edge** 	**SoC/CPU** 
Acceleration Methodology	**PCIe offload** 	**Platform-level** 	**Micro-architectural** 
Energy efficient System	**Multi-threaded EC** 	**Multi-DPU** 	**RVV Multi-threading** 

High-performance Computing and Big Data Processing

[P8] **Safe Hardware Multi-threading for Erasure Coding in Distributed File Systems**

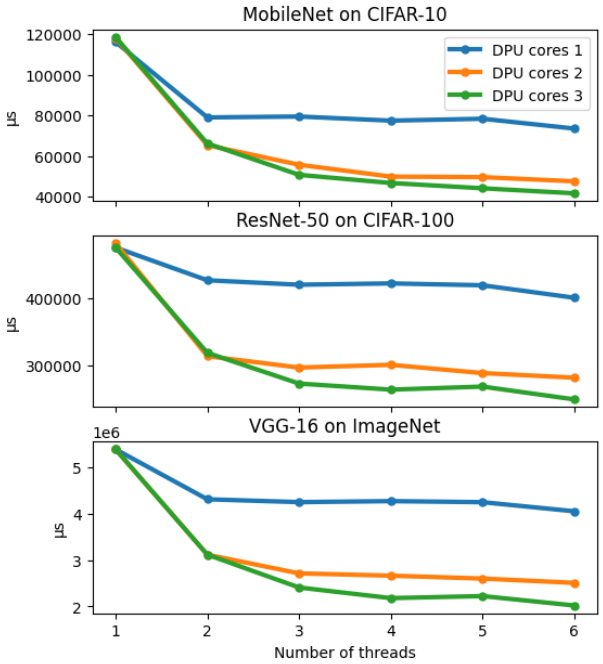
- Energy efficient EC acceleration framework
- VF Proxy: thread-isolated middleware integration
- Virtual accelerated cluster

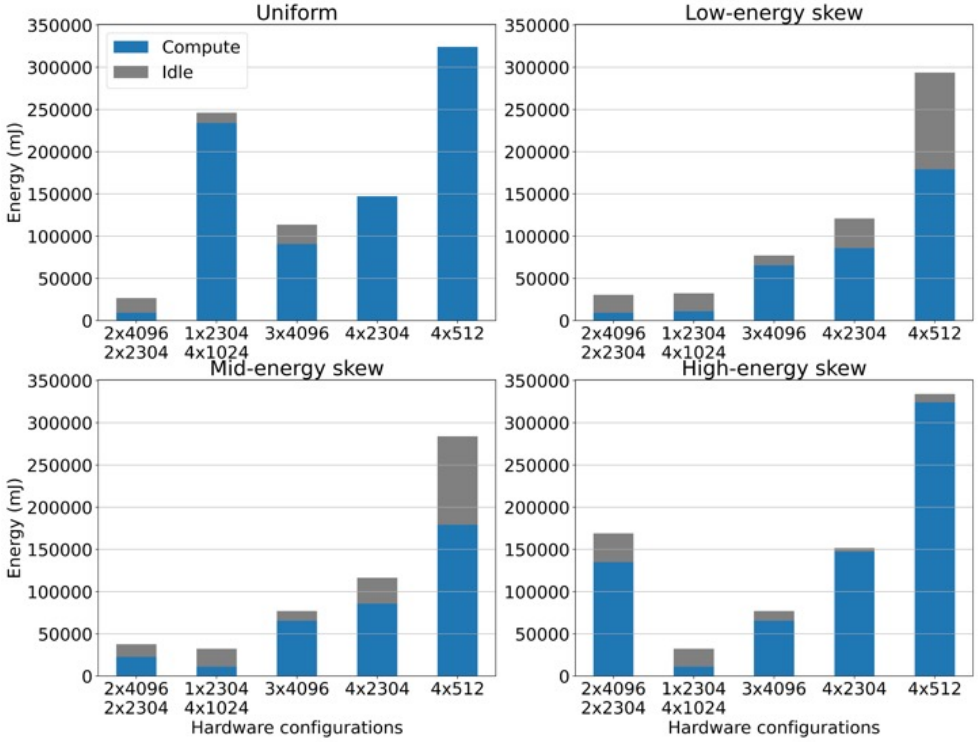


Edge Computing and Deep Learning

Multi-DPU: Heterogeneity for Multi-tenant Edge-AI

- [P3] Multi-threaded and multi-tenant MPSoC evaluation
- [P7] Enable energy efficient multi-tenant edge-AI

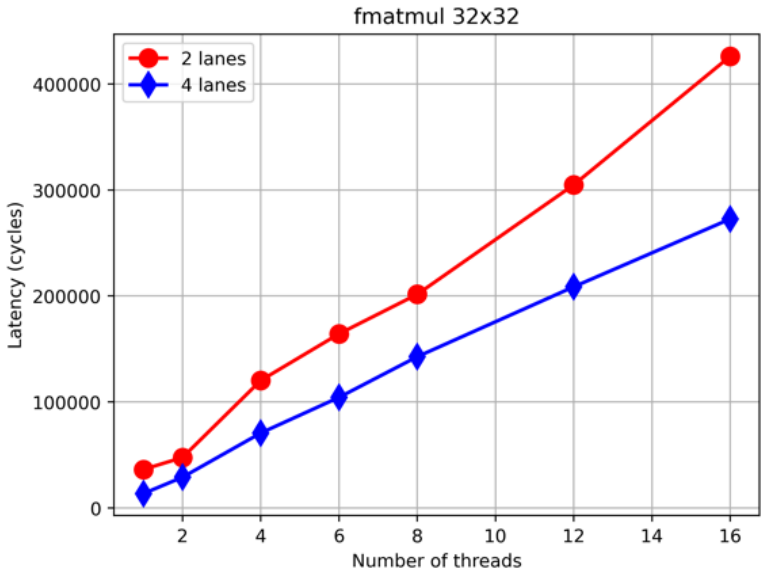




Low-power SoCs and ISA Acceleration

RISC-V Vectors and multi-threading:

- [P1] V-extension micro-architectural design
- [P4] OS support for open-source vector co-processor
- [P9] RVV multi-threading evaluation



Conclusions

- This thesis presents and validates the application of **cross-domain** approaches for:
  - **Hardware consolidation**
  - **Energy efficiency**
- We aim to **empower computer architects** to design energy efficient digital computing systems
  - Across the computing landscape
  - Without relying on technologies
- Targeted digital computing domains:
  - HPC/Cloud
  - Edge computing
  - SoC/CPU