

Saba Ebrahimi

 sabaebrahimi |  saba.eb18@gmail.com |  Scholar

Research interests: Computer architecture, processing-in-memory, OS/kernel-level systems for distributed and heterogeneous computing.

EDUCATION

- 2024 - present MSc - Computer Engineering, Computer Architecture at **Shahid Beheshti University**
Hardware Accelerators and High Performance Computing lab (GPA: 18.29/20)
- 2019 - 2024 BSc - Computer Engineering at **Shahid Beheshti University** (GPA: 18.31/20)
Ranked among top 5 students — Head Teacher Assistant of *Operating Systems, Advanced Programming, Embedded systems*.
- 2016 - 2019 Diploma - Mathematics and physics at **Farzanegan1 High school** (GPA: 19.55/20.0)

PUBLICATIONS

1. A. Noohi, M. Derispour*, **S. Ebrahimi*** and A. Barbalace, "NTSM: OS-Level Memory Middleware for Scaling AI Without Boundaries," 27th ACM/IFIP International Middleware Conference, Universitat Rovira i Virgili - Spain, 2026
2. M. Aghaei*, **S. Ebrahimi***, M. S. Arafati, E. Cheshmikhani, D. Rahmati, S. Gorgin and J. Kim, "Modeling and Simulation Frameworks for Processing-in-Memory Architectures," *arXiv preprint arXiv:2512.00096*, 2025. DOI: 10.48550/arXiv.2512.00096.

SKILLS

- **Systems & Architecture:** C, C++ - Linux Kernel - Embedded development - Cuda - IoT - Distributed Systems - RTOS - AUTOSAR - Verilog & FPGA
- **Others:** Python - llama.cpp, vllm, OpenVino - Java Script and Web - Git, Jira

RESEARCH

OS Research Assistant — The University of Edinburgh — United Kingdom Oct 2024 - Jan 2026

Supervised by Dr. Antonio Barbalace (system-nuts lab). Worked on over-network thread migration for load balancing across distributed nodes via Linux kernel modification.

- Implemented kernel-level thread migration and startup by modifying the scheduler, enabling seamless workload continuation across nodes.
- Extended the kernel page-cache mechanism to support a shared file system across multiple machines.
- Co-authored resulting paper (NTSM, Middleware 2026): $4.4\times$ inference TTFT and $6.5\times$ distributed training memory reduction on a 34-node RDMA cluster.

AI Hardware R&D — Algonet — Iran

Feb 2026 - present

- Optimized LLM inference for edge devices using llama.cpp, ik_llama, and custom inference runtimes.
- Improved inference performance by achieving $2\times+$ higher prefill throughput and 50% faster decoding through memory optimization, cache-aware execution, multi-threading, and pipeline parallelism.
- Designed heterogeneous inference pipelines utilizing Intel Core Ultra NPUs and Intel GPUs, including an OpenVINO runtime supporting multi-request GPU–NPU execution.
- Implemented hardware-aware inference optimizations including MoE execution, KV-cache optimization, quantization, speculative decoding, and Flash Attention.
- Built automated benchmarking and profiling infrastructure to measure latency, throughput, and memory footprint across Edge devices (Nvidia Jetson, Apple M series,...).

Research Assistant — Shahid Beheshti University

Feb 2025 - present

- Conducted research in the Hardware Accelerators Laboratory, focusing on Processing-in-Memory (PIM).
- Implementing master's thesis on lookup table (LUT)-based PIM techniques, optimizing Multi-Head Attention (MHA) and Grouped Query Attention (GQA) for LLM inference.
- Developed a LUT-based PIM simulator for the UPMEM, extending the uPIMulator framework.

Research Assistant — Shahid Beheshti University — [GitHub](#)

Aug 2023 - Mar 2024

- Led research focusing on developing a **Self-Adjusting Thermostat** that determines user's thermal comfort range.
- Developing an on-device reinforcement learning algorithm based on occupant's schedule.
- Adding the edge device to connect multiple thermostats and learn based on online information. Finally, putting them all together using an IoT platform.

Research Assistant — RADIANT IOT lab of Sharif University — Iran

Nov 2022 - Dec 2023

- Led research initiatives at the RADIANT IOT Lab, focusing on optimizing parent selection algorithms for the RPL protocol on Contiki OS.
- Explored IOT communication protocols, localization algorithms and contributing to a comprehensive understanding of cutting-edge technologies.

WORK EXPERIENCE

Embedded Developer — Software Motion Engineering Services — China

July 2023 - May 2024

- Contributed in software development projects for a Chinese company specializing in autonomous driving (ADAS) in the automotive sector.
- Working extensively with AUTOSAR architecture, particularly focusing on OS and Flash modules.
- Deploying and optimizing code on AURIX boards to improve system functionality and performance.

Full Stack Developer — Mohaymen ICT — Iran

Apr 2021 - Feb 2023

- Worked as a front-end developer on a major financial project, building the user interface with React.
- Contributed to back-end development using SpringBoot, Kafka, and Docker for smooth integration.
- Contributed in the front-end development for an innovative data analysis product, using Angular.

PROJECTS

Controller for Maxon EPOS Driver

[GitHub](#)

Working with Raspberry Pi 3B, launched Linux-based desktop application with PyQt, made connection between Raspberry Pi and Maxon EPOS driver.

Desktop Application for Smart Firefighter suit.

[GitHub](#)

This project was an application developed with Electron JS, that receives firefighter's vital signs from a smart suit. Data is transferred in Bluetooth and serial protocol.

DTMF Signals Decoder

[GitHub](#)

Designing DTMF Signal Decoder with STM32F401RE, using HAL library

REFERENCES

Dr. Antonio Barbalace — University of Edinburgh

abarbala@ed.ac.uk

Dr. Dara Rahmati — Shahid Beheshti University

d_rahmati@sbu.ac.ir

Dr. Hossein Attarzadeh — Shahid Beheshti University

h_attarzadeh@sbu.ac.ir