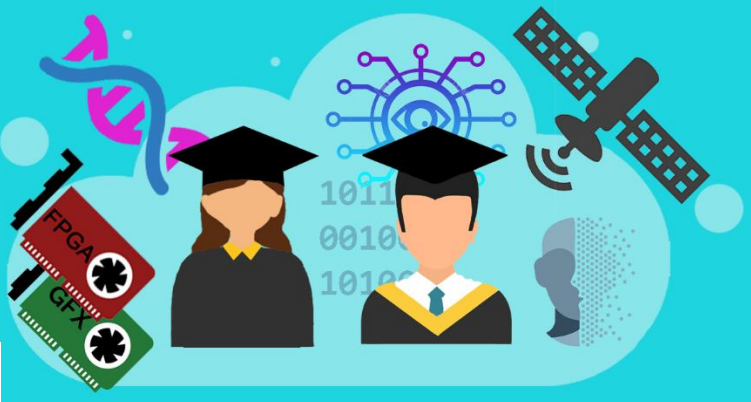
Diploma Thesis

Microproccessors and Digital Systems Laboratory

&

ETH zürich



Hardware–Software–OS Co-Design for Efficient & Reliable Physical AI

Co-designing the OS, hardware, and runtime is a promising path to make Physical AI (robotics, embodied agents, autonomous systems) faster, more predictable, and safer. In this thesis, we will focus on practical mechanisms, including scheduling, memory, I/O, and accelerator management. These mechanisms are crucial for minimizing tail latency, bounding jitter, reducing energy consumption, and improving reliability for physical AI workloads.

1. Memory-Centric Computing Architectures for Physical AI

This research thrust focuses on designing new memory and storage architectures to overcome the data movement bottleneck that limits the performance and energy efficiency of Physical AI systems. We will investigate Processing-in-Memory (PiM) and In-Storage Processing (ISP) paradigms, which embed compute capabilities directly within or near the memory chips or the storage medium[1,2]. This approach enables massively parallel operations on data where it resides, thereby fundamentally reducing the latency and energy costs associated with moving large datasets. Our work will involve profiling real-time physical AI workloads, such as Visual-SLAM [3] and Vision-Language-Action models [4], to identify their unique memory access patterns. Based on these insights, we will design and evaluate specialized hardware, memory controllers, and the necessary system software to enable these next-generation, data-centric architectures. The goal is to create systems that are not only faster but also more predictable and reliable for real-time robotic and embodied agent applications.

2. Memory hierarchy optimizations for real-time guarantees

This thrust focuses on making sure the AI's "brain" has immediate access to the data it needs. We are going to develop ways to manage memory so that critical, real-time tasks, like avoiding an obstacle, are never delayed by slower, less important tasks.
Our approach will involve changes both (i) at the CPU/GPU/NPU microarchitecture level and (ii) at the main memory controller.

3. Leveraging physical feedback for scheduling & resource control

This research thrust is about teaching a system to learn from the successes and failures it is responsible

for. We will be creating a system that leverages real-world feedback—like whether the underlying robot successfully picked something up—to adjust its decisions in real time.

4. Scheduling across the device–edge–cloud continuum

This work is about intelligently deciding where different parts of an AI task—like perception or planning—should run [5]. Our goal is to design a traffic controller for computing, ensuring that the right tasks are executed on the appropriate device, whether it's the robot itself, a nearby server, or a remote cloud, to ensure everything runs smoothly and efficiently.

5. Infrastructure for hardware-in-the-loop (HIL) robotics simulations

This thrust focuses on creating a high-fidelity test environment for robots. We will be building sophisticated simulators that can mimic real-world conditions, including potential faults and interferences, to test how a robot will perform before it's deployed. This allows for rigorous testing in a safe and controlled setting. We will explore ideas presented in one of our latest works on simulation [6].

6. System design for swarm robotics, including connectivity workloads

This research direction is all about helping groups of robots work together seamlessly. We will be designing the communication systems (both HW and OS/SW) that allow multiple robots to share information, coordinate tasks, and adapt to changing conditions, even if their network connection is unreliable. This enables them to accomplish complex missions as a single, cohesive unit.

Requirements:

1. Strong coding (mainly C/C++) skills
2. Strong computer architecture and operating systems background
3. Strong work ethic

References:

[1] Onur Mutlu et al. "Memory-Centric Computing: Recent Advances in Processing-in-DRAM"

[2] Onur Mutlu et al. , "Memory-Centric Computing: Solving Computing's Memory Problem"

[3] Visual SLAM, https://github.com/NVIDIA-ISAAC-ROS/isaac_ros_visual_slam

[4] pi0 Model, https://www.physicalintelligence.company/download/pi0.pdf

[5] Cog et al., "D3: A Dynamic Deadline-Driven Approach for Building Autonomous Vehicles" Eurosys 2022

[6] Kanellopoulos et al., "Virtuoso: Enabling Fast and Accurate Virtual Memory Research via an Imitation-based Operating System Simulation Methodology."

If you are interested, please email:

Konstantinos Kanellopoulos, konkanello@gmail.com, https://konkanello.github.io/
Dimosthenis Masouros Ph.D.: (dmasouros@microlab.ntua.gr)
Prof. Onur Mutlu, omutlu@gmail.com, https://people.inf.ethz.ch/omutlu/
Prof. Dimitrios Soudris: (dsoudris@microlab.ntua.gr)