Nikoleta Markela Iliakopoulou

email: nmi4@illinois.edu
Linkedin: nikoleta-iliakopoulou
Mobile: (+1) 217-904-2582

Research Interests

My research interests lie in the areas of Computer Architecture, Cloud Computing and Machine Learning. I investigate ways to make cloud environments for machine learning efficient and reliable from an architecture and systems perspective. My work focuses on optimizing large language model (LLM) inference serving, including full system design, from workload scheduling in data centers to GPU hardware design.

Education

University of Illinois at Urbana-Champaign *2022–Present*

Doctor of Philosophy (PhD) student in Computer Science (CS)

- *Advisor:* Josep Torrellas
- *Expected Graduation:* May 2027
- *Passed Qualifying Exam:* Spring 2024
- *Selected Advanced Courses:*
  - *Parallel Computer Architectures* CS-533
  - *Advanced Operating Systems* CS-523
  - *High Speed and Programmable Networks* CS-598HPN
  - *Architectures for Mobile and Edge Computing* CS-534

National Technical University of Athens *2016–2022*

Diploma (5-year joint degree) Electrical & Computer Engineering (ECE)

- *Major:* Computer Science
- *Thesis:* “Virtualization Techniques on Embedded Systems with further application on Satellites”. *Advisor:* Dimitrios Soudris.
- *Selected Courses:*
  - *Advanced Topics in Computer Architecture* ECE-3352
  - *Distributed Systems* ECE-3377
  - *Parallel Processing Systems* ECE-3257
  - *Operating Systems Laboratory* ECE-3237
  - *Embedded System Design* ECE-3361

Publications

N. Iliakopoulou, J. Stojkovic, C. Alverti, T. Xu, H. Franke, and J. Torrellas, “Chameleon: Adaptive Caching and Scheduling for Many- Adapter LLM Inference Environments.” https://arxiv.org/abs/2411.17741, 2024.

J. Stojkovic, C. Alverti, A. Andrade, **N. Iliakopoulou**, T. Xu, H. Franke, J. Torrellas. (March 2025). ”Concord: Rethinking Distributed Coherence for Software Caches in Serverless Environments”. In Proceedings of the 31st International Symposium on High-Performance Computer Architecture (HPCA).

J. Stojkovic, **N. Iliakopoulou**, T. Xu, H. Franke, J. Torrellas. (June 2024). ”EcoFaaS: Rethinking the Design of Serverless Environments for Energy Efficiency ”. In Proceedings of the 51th International Symposium on Computer Architecture (ISCA).

Research Experience

Research Intern, Network Research Group, NVIDIA *05/2024 - 08/2024*

Advisor: Dr Nic McDonald, Exhaustive GEMM profiling for architectural analysis and exploration, Santa Clara, CA.

Research Visiting Scholar, Hybrid Cloud, IBM Research *05/2023 - 08/2023*
Advisor: Dr Hubertus Franke, Towards a high performant and resource efficient platform for Machine Learning as a Service (MLaaS), Thomas J. Watson Research Center, NY.

Research Assistant at the University of Illinois at Urbana-Champaign *08/2022 - Present*
Advisor: Prof. Josep Torrellas, Building hardware-software LLM inference stack, Urbana-Champaign, IL.

Research Intern at OHB-Hellas *06/2021 - 07/2021*
Advisor: Dr Mathieu Bernou, Hardware Virtualization for Mixed-Criticality software applications, Athens, GR.

School of Electrical and Computer Engineering, National Technical University of Athens. *04/2021 - 07/2022*
Advisor: Prof. Dimitrios Soudris, Virtualization Techniques on Embedded Systems with further application on Satellites.

Projects

Efficient Many-Adapted LLM Serving (UIUC). Large language models (LLMs) are becoming integral to numerous applications, transforming interactions with technology, content generation, and data analysis. In this project, I examined the computational and system challenges of deploying generative LLMs using the "pretrain-then-finetune" approach. Despite its potential for efficient inference, current practices suffer from inefficiencies that significantly degrade the latency and throughput of such environments. To solve this, I designed an adapter-aware system that proposes a set of techniques for smart request scheduling, adapter caching, and prefetching.

High Performant and Resource Efficient ML as a Service (IBM). I investigated Ray, an emerging, open-source, unified compute framework for distributed machine learning (ML). As it was initially designed for long running tasks, i.e., ML training, the argument arose as to whether it could run short ML inference workloads efficiently. I performed a thorough characterization of ML inference workloads using open-source benchmarks. I identified and analysed Ray's overheads in the scope of ML inference in the cloud. I proposed an alternative implementation by effectively sharing memory state across requests of the same ML workloads. Instead of using Ray's expensive worker-process pool, the scheduler can bypass it and send requests to execute on a lightweight process-pool that shares read-only libraries and model weights, by effectively exploiting the fork technique. I built a prototype on top of Ray that leads to up to 3 times less memory consumption while maintaining high performance.

Model Parallelism Techniques for Inference Serving using SmartNICs (UIUC). Large ML models are resource-hungry and bring great challenges in achieving good performance and meeting service-level agreement. Fortunately, techniques like model partitioning can greatly reduce request latency by leveraging task and data parallelism. I explored model partitioning techniques using new hardware like SmartNICs that offer an option to offload tasks from host CPUs to "cheaper" cores, while saving expensive CPU cycles for other resource-hungry workloads. By exploiting the heterogeneous nature in ML models and offloading the non compute-intensive parts, median and tail latency can effectively be reduced under high load. I built a prototype system with Nvidia Bluefield SmartNICs, which yields a median latency decrease of up to 42.7% and tail latency decrease of up to 72.5%.

Virtualization Techniques on Embedded Systems with further application on Satellites (NTUA). Recent growth in the space industry has drawn attention to Satellite as a Service (SaaS). The primary goal of SaaS is to maximize the use of orbital resources while introducing novel concepts, such as the idea of data processing onboard a satellite. I investigated different virtualization strategies for mixed-criticality software applications and examined and compared metrics such us isolation, real-time performance, fault tolerance, flexibility and ease of programming. I built a prototype on top of Jailhouse Hypervisor, combining static-partitioning bare-metal virtualization with containerization, deployed on top of an ARM single-board computer.

OHB-Hellas I investigated different virtualized environments to handle securely, abstractly and efficiently different types of AI/ML based software applications. I simulated the communication chain from the end-to-end user to the satellite processing board and characterized the network overheads, while having a (non-)persistent link. I also built and tested aarch64 Docker images for different ML applications.

Technical Skills

Programming Languages	Python, C/C++, Java, P4, JavaScript, Assembly (mips, 8085, 8086, AVR)
Hardware Description Languages	VHDL, Verilog
Operating Systems	Linux, MacOS, Windows
Container Management	Docker, Kubernetes
Databases	MySQL, MongoDB
Version Control	Git/Github/Gitlab
Frameworks & Tools	Ray, Pytorch, CUDA, Tensorflow, NVIDIA DOCA, OpenMP, MPI, NumPy, SciPy
SoC Development & HLS	Vivado Design Suite

Honors & Awards

Sensors and Coding Hackathon 2018 *2018*

- 24hour Hackathon organized by IEEE NTUA S.B. – ADVANTAGE AUSTRIA – SICK Sensors Intelligence
- 1st Place as member of the “Sensates” team

Eestech Challenge Competition 2018 *2018*

- European Annual Programming Competition. The year’s topic: “Big Data”
- 1st Place in local round held in Athens, Greece
- 3rd place in final round held in Novi Sad, Serbia

Languages

Greek (native), English (professional), French (fluent), German (elementary)