Selin Yıldırım

PhD Student in Computer Science

University of Illinois at Urbana-Champaign

+1447902-2251 | seliny2@illinois.edu | /selinnilesy | selinnilesy

Summary

4th year PhD student in Computer Science with a passion in producing impactful academic works at the intersection of High Performance Computing and Generative AI. Taking a role in a brilliant academic research group where cutting-edge innovations and team collaboration are basic priorities. An excellent communicator and team player with the ability to meet deadlines and quickly resolve problems.

Experience

- RESEARCH ASSOCIATE - EFFICIENT GENAI

San Jose, California, USA

Institution: Advanced Micro Devices

05/2025 - Current

Speculative Decoding for Image Generation: Exploring accelerated image generation on continuous and discretized image token idomains with novel, fast and efficient speculative decoding algorithms. Achieved impressive speedup up to 3.6x over vanilla image generation.

- RESEARCH INTERN - GENAI

Remote

Institution: NVIDIA

08/2024 - 12/2024

Towards Memory-Efficient Decoding on Constrained GPUs: Developed an hardware-aware, memory-efficient and fast speculative decoding approach that leverages training-free tree pruning techniques on tree-based speculative decoding, specifically for inferencing on constrained GPUs. *SpecMemo: Speculative Decoding is in Your Pocket.*

- DEEP LEARNING PERFORMANCE INFERENCE INTERN

Santa Clara, California, USA

Institution: NVIDIA

05/2024 - 08/24

TensorRT-WinAi Helped Nvidia develop smarter deep learning compiler algorithms within TensorRT inference framework. Built an extraordinary network of fellow researchers and engineers. Worked on a novel training-free KV cache compression algorithm for LLMs.

- GRADUATE RESEARCH ASSISTANT, Coordinated Science Laboratory

Champaign, IL, USA

Institution: University of Illinois at Urbana-Champaign

05/2023 - Current

Faster Decoding: Improving sparsity in the speculative heads of *Medusa* to make inference faster and more efficient : *Paper*.

GPU Concurrency: Provided a resource efficient framework for deploying mixed deep learning workloads on NVIDIA GPUs where full isolation of GPU resources are provided to cloud multi-tenants. Working hands-on with hypervisors, Nvidia MPS, Nvidia MIG, CUDA, Kubernetes and Docker.

- GRADUATE RESEARCH ASSISTANT, Parasol Laboratory

Champaign, IL, USA

Institution: University of Illinois at Urbana-Champaign

08/2022 - 05/2023

Exoplanet: Achieved remarkable speedups for the scientific code in Exoplanet Project on distributed GPUs, funded by NASA. Worked on parallel graph traversal problems such as bottom-up BFS with direction optimization and finding dynamic scheduling strategies of threads and data distribution, as well as investigating different data decomposition techniques for memory bound problems.

- UNDERGRADUATE RESEARCH ASSISTANT

Ankara, TURKEY

Institution: Department of Computer Engineering, Middle East Technical University

10/2021 - 08/2022

Job Description: SIAM AN'24 contributed talk on paper:
PARS3: Parallel Sparse Skew-Symmetric Matrix-Vector Multiplication with Reverse Cuthill-McKee Reordering , where sparsity in Skew-Symmetric Sparse Matrix-Vector Multiplication problem is address efficiently. This paper has been ranked the 2nd in the undergraduate research competition at Middle East Technical University.

- HIGH PERFORMANCE COMPUTING TRAINEE, ERASMUS+ TRAINEESHIP

Stuttgart, Baden-Württemberg, GERMANY

Institution: High Performance Computing Stuttgart (HLRS), University of Stuttgart

07/2021 - 10/2021

Job Description: Worked on European Union's Horizon 2020 research and innovation program project, namely *ExaHyPE*, which aims to forecast and simulate possible natural disasters to minimize the damage by better predicting them. Worked with the SPMT Team, to optimize intermediate I/O operations in the kernel by using parallel programming skills. Developed paralell Sieve of Eratosthenes Prime Finder Algorithm with Intel TBB. Asynchronous Tasks Optimization on I/O Operations of ExaHyPE's Compressible Navier-Stokes Example.

- COMPUTER VISION, C++ DEVELOPER *Ankara, TURKEY*

COMPANY: ARGOS AI, MIDDLE EAST TECHNICAL UNIVERSITY TECHNOPARK *10/2020 - 3/2021*

Job Description: C++ developments to control airport cameras and capture point of views, then by processing those images for foreign object detection to keep runways and aprons of airports safe for upcoming flights. Worked on isolation of a microservice from a big monolith C++ project, which controls a remote Pan-Tilt Unit and a Vimba camera, then automated the microservice by deploying it on a gRPC server.

- MACHINE LEARNING INTERN *Ankara, TURKEY*

COMPANY: OTSIMO, MIDDLE EAST TECHNICAL UNIVERSITY TECHNOPARK *06/2020-10/2020*

Job Description: Worked on Speech Recognition neural network whose development is inspired by Apple Siri's algorithm, on remote computing platforms and performed data analysis using Big Query on Google Cloud. Training and testing ML models, supervised learning, adopting accuracy/loss calculations to trainings; implementing CLI for admin web interface using Golang.

Skill Set

Advanced Parallel Programming & GPU Performance Optimizations: Expertise in hand-written CUDA kernels, MPI, OpenMP, STAPL, Intel TBB for high-performance computing.

Generative AI & Machine Learning: In-depth experience with deep learning frameworks and LLM performance. Optimization of Transformers architecture and attention layers to reduce decoding latency and memory overhead.

Systems Architecture & Compiler Engineering: Proficient in C++, Python, Java, Assembly, ANSI C, Golang, Haskell, Prolog, with a focus on performance-critical applications. Skilled in compiler construction, LLVM, TensorRT and TensorRT-LLM for efficient model execution and deployment. Advanced understanding of computer organization, instruction sets, and low-level architecture (X86, Y86) optimization.

System Performance Analysis & Tuning: Experience with Nsight Systems, Compute, and nvidia-smi for in-depth performance profiling and tuning. Strong background in low-level instruction cycle analysis and performance improvements in C/C++ environments.

Education

University of Illinois at Urbana-Champaign *IL, USA*

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE *2022 - Current*

Compiler, Architecture and Parallel Computing research track. **GPA:3.9**

Middle East Technical University *Ankara, TURKEY*

BACHELOR OF SCIENCE IN COMPUTER SCIENCE *2017 - 2022*

Top ranked CS department with the highest research output in Turkey. Top-interests were Parallel Programming, Guided Research, Computer Architecture, Operating Systems and Quantum Computing.

Teaching

University of Illinois at Urbana-Champaign *IL, USA*

CS521 MACHINE LEARNING AND COMPILERS *8/2025 - Current*

Siebel School of Computing and Data Science

Publications and Conferences

10/2025	**ICCAD'25**, A3C3 – AI Algorithm & Accelerator Co-design, Co-search, and Co-generation	*Munich, Germany*
07/2024	**SIAM'24**, *Contributed talk, PARS3: Parallel Sparse Skew-Symmetric Matrix-Vector Multiplication with Reverse Cuthill-McKee Reordering*	*Washington, USA*
09/2021	**Euro MPI 2021**, Attended as an audience guest of HLRS to one of the most important conferences in High Performance Computing in Europe, Euro MPI, which hosts a professional international committee and is conducted by Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities, Germany.	*Munich, Germany*

Projects

10/2021 - 07/2022	**Wanderlust**, Senior project for the Computer Engineering Design Course, CENG491/2. Wanderlust is a mobile application planning holistic touristic trips for people. It has a new multi-variate Traveling Salesman algorithm. In the core, AI-based combinatorics algorithm tries to calculate the maximum satisfaction all-inclusive path, considering user inputs, accommodation preferences, transportation tickets, time, and the city characteristics that might affect derived satisfaction.	*Department of Computer Engineering, METU*
07/2021	**Parallel Finding Prime Numbers**, Conducted as a summer internship project, *Parallel Finding Primes*. Deployed on Slurm HPC platform in C++ and MPI. Researched optimization methods on Sieve Algorithm of Eratosthenes and tried to implement methods proposed in *Parallel Prime Sieve: Finding Prime Numbers* and analyzed performance.	*Department of Computer Engineering, METU*

Certifications

10/2020	**Machine Learning by Andrew NG, Stanford University, Grade:98.82/100**, *Certificate Credentials*	*Coursera*
09/2020	**Problem Solving**, *Certificate Credentials*	*Hackerrank*
09/2020	**C++**, *Certificate Credentials*	*Hackerrank*
09/2020	**Intermediate C**, *Certificate Credentials*	*Hackerrank*
11/2019	**Extracurricular : Work and Travel Exchange Program**, Photographer and Sales Assistant , *Certificate Credentials*	*Maryland, the USA*

Achievements

University

07/2022	***Second Best Research Award***, Guided Research, Department of Computer Engineering, Middle East Technical University	*Ankara, TURKEY*
2022-2023	**Departmental Fellowship**, University of Illinois at Urbana-Champaign	*IL, USA*
2017-2022	**Non-refundable Grant Award for Skillful Undergraduate Students**, *Sema Yazar Youth Foundation*	*Ankara, TURKEY*
5x	**Placed in High Honor Students Roll**, Middle East Technical University	*Ankara, TURKEY*

Exams

2017	**Top 0.05% among 2.162.895 participants**, National University Entrance Exam (LYS)	*Turkey*
2013	**Top 0.1% among 1.112.604 participants**, National High School Entrance Examination (SBS)	*Turkey*

Languages

English	IELTS: 7.5 (C1) , European Online Linguistic Support: C1
German	A2
Latin	A1
Turkish	Native Language