Erel Ozen

Irvine, CA
949-339-7540 erelozen@gmail.com linkedin.com/in/erel-ozen eozen1.github.io

Education

Georgia Institute of Technology **May 2027**
B.S/M.S Computer Science (Intelligence & System Architecture) — ***GPA: 4.00*** *Atlanta, GA*

Relevant Coursework: Design of Operating Systems, Systems & Networks, Design & Analysis of Algorithms, Data Structures & Algorithms, High Performance Computing, Compilers, Machine Learning, Computer Organizations & Programming (*Teaching Assistant*), Artificial Intelligence, Perception & Robotics, Object-Oriented-Programming

Experience

Apple **May 2025 – August 2025**
Software Engineering Intern | CoreOS, Virtualization, MCP *Cupertino, CA*

- Architected and deployed a multi-agent AI triage tool for **1600+** engineers, automating root-cause analysis and reducing manual triage time by 70%.
- Developed a scalable backend service using **MCP** servers to orchestrate AI agents that autonomously analyze large logs, query internal ticketing systems, and interface with debugging tools to generate actionable debugging steps.
- Integrated the platform into Apple's CI framework using webhooks to automatically analyze nightly test failures.
- Resolved critical bugs in high-performance virtual machine (VM) network drivers by re-engineering state management logic, significantly improving network reliability and system performance at scale.
- Selected as **one of 10** interns company-wide to present to Apple's SVP of Software Engineering, Craig Federighi.

OKSI **May 2024 – May 2025**
Software Engineering Intern | Machine Learning, LoRA, AWQ, NATS.io *Torrance, CA*

- Secured a **$1.5M** contract renewal from DARPA after presenting an end-to-end AI automation platform.
- Deployed Llama-3 on edge devices with **400%** less VRAM by leveraging activation aware weight quantization.
- Developed a full-stack React application featuring an NLP pipeline with LoRA fine-tuning, NATS.io, & Synadia Cloud.
- Architected a PDF form processing microservice that reduced manual entry with an auto-populating question-wizard.
- Developed GNC algorithms for a drone with path planning and SLAM in ROS2 for real-world autonomous navigation.

OKSI **June 2023 – August 2023**
Software Engineering Intern | Embedded Systems, SLAM, IMU, GPS *Torrance, CA*

- Orchestrated configuration and calibration of multispectral cameras, GPS, and IMU for machine-learning applications.
- Evaluated camera drivers on a Google Coral Board, implementing debayering techniques for optimal performance.
- Developed high-speed serial communication between microcontrollers and avionics, focusing on data pipeline efficiency.
- Implemented simultaneous localization and mapping (SLAM) using LiDAR, IMU, and GPS for autonomous navigation.

Projects

Vidur Autoscaling | *LLM Inference Simulator Extension* **Fall 2025**

- Implemented reactive autoscaling policies for distributed LLM inference using traffic analysis and throughput estimation
- Optimized O(1) traffic envelope algorithms for real-time token arrival rate tracking in production systems
- Developed LOR scheduler with autoscaling awareness to minimize latency during replica scale-down operations
- Engineered custom autoscaling policy generating Pareto-optimal latency-cost trade-offs across service levels

UNIX Kernel Development | *OS development for a UNIX v6 Kernel* **Spring 2025**

- Implemented core OS components for a UNIX V6-based teaching OS in ANSI C on x86, including the bootloader, virtual memory management, file system, and CPU scheduling
- Developed system-level functionality such as interrupts, concurrency mechanisms, user and kernel threading, and basic networking, focusing on process execution, resource management, and inter-process communication

Organizations

Systems for AI Lab @ Georgia Institute of Technology **May 2025 – Current**
Research Assistant *Atlanta, GA*

- Building LLM-serving distributed system and adding support for various types of attention (DeepSeek, Qwen)
- Increased LLM throughput by optimizing prefix caching strategies through dynamic memory management
- Extended core functionality of PyTorch to enable custom memory allocation, using C++, CUDA, and Docker

Technical Skills

Languages: Python, C/C++, Java, Go, JavaScript, TypeScript, Bash, SQL, x86 Assembly
Frameworks: TensorFlow, PyTorch, ROS/ROS2, OpenCV, CUDA, LLVM, Triton, CUTLASS, JAX, XLA, Kubernetes, Matplotlib, ReactJS, React Native, NATS.io, JUnits, pySerial, HF Transformers, vLLM Node.js, Express.js, pytest
Technologies: RESTful APIs, AWS, Git, Docker, GenAI, LLM Inference/Serving, Quantization, Autoscaling, Distributed Systems, Virtualization, CI/CD, Concurrency, SLAM, Controller Area Network (CAN), Micro Controllers, LoRA