

Vaibhav Mattoo

217-493-8071 | [Website](#) | [Github](#) | vmattoo2@illinois.edu

EDUCATION

University of Illinois Urbana-Champaign

Bachelor of Science in Computer Engineering

GPA : 3.96

Aug 2023 - May 2026

Relevant Coursework: Distributed Systems, Operating Systems, Computer Security, Networked IoT, Machine Learning, Deep Generative Models, Computer Architecture, Data Structures and Algorithms

EXPERIENCE

Blockchain Research and Development intern

October 2025 – Present

Kocree Inc.

- Designed a Proof of Useful Work and Proof of Stake based L1 blockchain for IP enforcement and fairness in information lattice-based creative generation platform. Developed cryptographic primitive for "creativity" computation based discrete log problems.
- Tentative white-paper due based on novel scheme for scalable on-chain IP enforcement and trade on custom architecture for Etherium Virtual Machines, with minimum consumption of network resources.

Distributed Inference Engineering Intern

May 2025 – August 2025

Tandemn Labs

- Designed and implemented efficient latency-optimized implementation of LLM request batching in Rust, efficiently utilizing heterogeneous infrastructure across geographically separate super computing clusters.
- Implemented networking stack for peer-to-peer communication of compute nodes across variable latency links between different data centers to minimize perceived response time for inference, based on Iroh with noise protocol.

Undergraduate Security Research - Information Trust Institute

August 2025 – Present

University of Illinois Urbana-Champaign

- Conducted in-depth threat modeling, and network measurements to find patterns in user activity and bandwidth consumption and designed a Tor routing protocol with ideal tradeoff between latency, bandwidth and anonymity.
- Designed an information theoretic anonymity metric based on a generic adversary model, which is decomposable into local anonymity terms and captures network diversity and possible traffic analysis vulnerabilities.

Undergraduate AI Research - Huimin Zhao Group

August 2025 – Present

University of Illinois Urbana-Champaign

- Designed and implemented a state-of-the-art model for regioselectivity prediction based on double cross-attention to capture enzyme-substrate interactions between a GNN substrate representation and ESM2 protein model.
- Optimized the framework to achieve 95.7%+ accuracy on regioselectivity prediction tasks in enzyme substrate reactions, due for publication in Nature journal.

NCSA SPIN AI Research Intern

August 2024 – May 2025

National Center for Supercomputing Applications

- Deployed APACE (AlphaFold on a distributed system with GPU acceleration) on Delta and Delta AI supercomputers. Set up systems for distributed task allocation using Ray and SLURM, dynamically allocated worker nodes across multiple GPUs and CPUs for AI fine-tuning and inference.
- Achieved 36% speedup in protein structure generation by implementing data compression techniques on amino acid string sequences. Made CPU optimizations by using IME to stage training database into SSD cache, to facilitate fast data tiering between compute nodes and file system in HPC environment.

Honors First-Year Design Lab Course Assistant

August 2024 – Present

University of Illinois Urbana-Champaign

- Provided mentorship to 100+ engineering students on conceptualizing, prototyping, and presenting circuit design and robotics projects. Assisted in teaching system-level design, common circuit design patterns and usage of lab equipment like multimeters and oscilloscopes.

Technical Director - Open Source at Illinois

Jan 2025 – Present

University of Illinois Urbana-Champaign

- Led workshops on Git/GitHub, Docker, Kubernetes, CI/CD for a student body of 1300 members.
- Led semester-long team projects, managing students groups to work on open source software projects.

PROJECTS

Bitchat-TUI | *Rust, Bluetooth*

- Implemented a terminal client with **5000+ users** for the BitChat protocol over Bluetooth Low Energy using the Noise Protocol Framework with X25519 ECDH key exchange, Ed25519 digital signatures, and AES-256-GCM authenticated encryption for secure peer-to-peer messaging.
- Developed cryptographic state machines for secure handshake establishment, ephemeral key generation, and message authentication using PBKDF2-SHA256 for channel password derivation with 100,000 rounds. Architected asynchronous Rust networking layer with Tokio runtime for concurrent BLE device discovery, connection management, and encrypted message fragmentation across multiple peer connections.

Alman: Context-Aware Shell Alias Manager | *Python, Rust, SQLite, Neural Networks, PCA*

- Designed an application with **1000+ users** based on a lightweight neural ranking model that prioritizes shell commands in real time using preceding command context, frequency, recency, length, and exit status to surface the most relevant commands to apply an alias for.
- Developed a system to intelligently suggest aliases for the identified commands using PCA to identify keywords and multiple alias suggestion schemes based on them.

SAVA Distributed Graph Processing System | *C++, TCP, Pregel Model, SDFS*

- Implemented SAVA on SDFS with a master-worker model, with active and standby masters for fault tolerance.
- Built a vertex-centric programming model exposing a Compute() function with message passing, edge traversal, and asynchronous execution through a Pregel superstep model. Outperformed GraphX by 170% (PageRank) and 400% (SSSP) via C++ implementation, random partitioning, and optimized TCP communication.

Holloway - Multi-Protocol Browser | *React Native, TypeScript, WebSockets, TLS*

- Developed a unified client for Gopher, Gemini, Finger, and Spartan protocols supporting concurrent multi-protocol connections via a modular React Native architecture.
- Implemented protocol-specific session state management, streaming parsers, and TLS (Gemini/Spartan) using native bindings, and ensured full cross-platform compatibility (iOS, Android, Linux). Integrated real-time media rendering, hypertext navigation, socket recovery and connection migration; decreased protocol handshake latencies with optimized socket pooling and reconnection logic.

IoT Self Driving Car | *Raspberry Pi, Python, OpenCV, TensorFlow, AWS IoT and Greengrass, PCB Design*

- Designed and programmed a Raspberry Pi-based autonomous vehicle prototype integrating sensors such as ultrasonic distance, velocity, servo motors, and camera for real-time environment sensing and control. Made a custom PCB in KiCAD to connect all hardware components effectively.
- Implemented mapping and navigation algorithms using ultrasonic sensor data and probabilistic SLAM concepts, enabling obstacle detection, avoidance, and route planning with A* for goal-directed pathfinding. Implemented MQTT based networking to enable communication with cloud based EC2 and S3 instance for data processing.
- Leveraged OpenCV and TensorFlow Lite to develop real-time object detection on low-power hardware, integrating traffic sign recognition to enhance vehicle decision-making and autonomous route control.

Custom Unix-Based Operating System | *C, RISC-V Assembly, QEMU, GDB, Git*

- Developed a Unix-like operating system from scratch, implementing core functionalities such as process management with support for forking and piping, virtual memory, ext2fs file system, ELF loader, locking, reference counting, support for 16 system calls and device drivers for a UART device, virtio disk and console.
- Designed and implemented a three-level page table structure for virtual memory management using the RISC-V Sv39 paging scheme, enabling user-mode memory isolation and efficient address translation. Implemented support for concurrent execution of multiple processes, each capable of running with up to 8 threads per process, scheduled based on a MLFQ scheduler.

Superscalar Out-of-Order RISC-V CPU | *SystemVerilog, Synopsis VCS, Verdi, Design Compiler*

- Created a speculative out of order RISC-V CPU with an ERR archiectyre, implementing RV32IM spec.
- Supports 1-8 issues/instruction commits per cycle, multiple integer execution units, variable size issue queues, etc. Split Load Store queue implemented with non-commited store forwarding and banked L1 cache. Optional parameters to change issue queue scheduling, TAGE based branch predictor, with a Full System UVM verification environment.

CUDA-Optimized Convolutional Layer for LeNet-5 | *CUDA C, Python, Nsight Systems, Nsight Compute*

- Developed forward-pass and backward-pass of convolutional and pooling layers in LeNet-5 architecture using CUDA C, improving performance for neural network inference.
- Optimized convolution by unrolling input feature matrices and using memory coalescing, tiling using shared memory and mask matrix storage in constant memory to reduce latency due to global memory access. Analyzed and fine-tuned the CUDA kernels using profiling tools like Nsight Systems and Nsight Compute, achieving significant performance improvements.

TECHNICAL SKILLS

Languages: Rust, TypeScript, JavaScript, SQL, C/C++, Python, Java, Golang, SystemVerilog, Dart, CUDA C

Tools: React Native, React.js, Node.js, Hadoop, Cassandra, AWS (EC2, S3, Lambda), Git, Docker, Makefiles, Linux

Libraries: Redux, WebSockets, PyTorch, TensorFlow, Scikit-learn, Ray, NumPy, Pandas