Minseob Shin

847-903-1319 minseob.shin11@gmail.com linkedin.com/in/minseob-shin minseobshin.vercel.app

Objective: Leverage expertise in hardware–software co-design to design and optimize high-performance computing systems

Education

University of Illinois at Urbana-Champaign **Aug. 2023 – May 2027**
Bachelor of Engineering in Computer Engineering GPA: **3.79/4.00**
Relevant Coursework: Computer Organization & Design (ECE411), Computer Systems Engineering (ECE391), Applied Parallel Programming (ECE408), Systems for GenAI (CS598), Digital Systems Laboratory (ECE385), Algs & Models of Comp (ECE374)

Projects

MeOoOw – Out-of-Order Execution Processor | *SystemVerilog, Computer Architecture & Microarchitecture* **Sept 2025 – Dec 2025**
- Architected an Out-of-Order (OoO) RISC-V processor using an explicit register renaming scheme with a Mapping Table and Free List to eliminate false data dependencies (WAR/WAW)
- Implemented a speculative execution pipeline featuring a Reorder Buffer (ROB) for in-order retirement and a Tournament Branch Predictor (Gshare/Local) to minimize control flow stalls
- Designed a high-performance memory hierarchy including a 4-way Set-Associative L1 Cache with Pseudo-LRU and a Split Load-Store Queue (LSQ) to handle memory disambiguation and forwarding
- Validated functional and timing correctness using the Synopsys toolchain (VCS/Verdi), performing extensive waveform analysis to resolve complex data hazards and ensure architectural state integrity

GPT-2 Inference Acceleration using CUDA | *CUDA, GPU Architecture & Acceleration, Deep Learning* **Sept 2025 – Dec 2025**
- Engineered custom high-performance CUDA kernels for the GPT-2 decoder pipeline, implementing Flash Attention-inspired tiling and KV-Caching to maximize SRAM reuse and eliminate redundant computations during autoregressive generation
- Developed high-throughput GEMM implementations by independently evaluating joint shared memory and register tiling against Tensor Core (WMMA API) acceleration to identify the optimal data-reuse strategy for transformer workloads
- Optimized kernel execution logic using tree-based parallel reductions and warp shuffles, effectively resolving synchronization bottlenecks and reducing latency in high-dimensional normalization layers
- Achieved a 1.86x end-to-end speedup (110ms to 59ms) solely through CUDA Streams to overlap memory transfers with compute; utilized NVIDIA Nsight Systems and Compute to identify memory-bound bottlenecks and reach hardware DRAM saturation

CosmOS - Operating System from scratch | *C, Operating Systems & Systems Programming* **Jul 2025 – Aug 2025**
- Developed a Unix-like operating system kernel for the RISC-V architecture, implementing core subsystems including preemptive multitasking, interrupt handling, and a robust system call interface
- Designed a Virtual Memory system (Sv39) featuring demand paging (lazy allocation) and page-level protection, ensuring complete process isolation and efficient physical memory management via multi-level page table walks
- Engineered an inode-based Read-Write Filesystem (KTFS) with indirect block addressing and a demand-paging cache layer to optimize disk I/O and support persistent storage
- Implemented VIRTIO block device drivers and a terminal interface, utilizing asynchronous I/O and ring buffers to facilitate high-throughput communication between the kernel and virtualized hardware

Experience

AMD **May 2026 - Dec 2026**
Incoming Software Engineer Co-op Austin, TX
- Will develop GPU profiling tools and optimize system performance within AMD's ROCm ecosystem, with involvement in ASIC bring-ups and cross-functional engineering collaboration

Future Architecture and System Technology for Scalable Computing **Jan 2026 - Present**
Undergraduate Research Assistant Urbana, IL
- Contribute to Retrieval-Augmented Generation (RAG) Acceleration using Compute Express Link (CXL) database and data compression

AMD-Xilinx Center of Excellence **Feb 2025 – Sept 2025**
Undergraduate Research Assistant Urbana, IL
- Contributed to the expansion of ScaleHLS, an MLIR-based high-level synthesis (HLS) framework, by integrating Dafny for formal verification to improve correctness and reliability in hardware design workflows
- Implemented computational kernels and convolutional neural networks (CNNs) directly in Dafny to explore verified hardware synthesis
- Collaborated with lab members through weekly meetings, slide presentations, and technical discussions to align research directions and communicate progress effectively

Leadership & Extracurricular

IEEE U of I **Mar 2024 – Present**
Web Developer Urbana, IL
* Led full-stack development of the official IEEE U of I Chapter website using Next.js, React, and TypeScript, creating a responsive platform that attracted 1,000+ visitors and enhanced event promotion, and community outreach
* Collaborated with club executives and members to continuously update and optimize the website, ensuring it meets the organization's evolving needs

Technical Skills

Languages: SystemVerilog, C, C++, CUDA, Python, Verilog, Assembly (RISC-V), Bash, JavaScript, Java, HTML/CSS
Frameworks & Tools: Git, Synopsys Toolchain, FPGA, Xilinx Vivado, Vitis, GDB, Nsight Toolchain, React