ABHISHEK SINGH

San Carlos, CA | +1 4846356147 | abhishek.singh199415@gmail.com
https://wordpress.lehigh.edu/abs218/

EDUCATION

Lehigh University — Bethlehem, PA
Doctor of Philosophy in Electrical and Computer Engineering, Computer Architecture Concentration — **May 2023**
GPA: 3.64/4

- Courses: Digital System Design, Memory System, Intro to VLSI, Physics of Semiconductor Devices

University of Rochester — Rochester, NY
Master of Science in Electrical and Computer Engineering, Computer Engineering Concentration — **May 2018**
GPA: 3.83/4

- Courses: Multiprocessor Architecture, Advanced Computer Architecture, Wireless Communication, Parallel Computing Using GPUs, Operating Systems, RF and Microwave Integrated Circuits, Wireless Sensor Networks

Mumbai University — Mumbai, India
Bachelor in Electronics Engineering
GPA: 3.5/4 — **June 2016**

- Selective Courses: Embedded System Design, Microcontrollers and Application, Microprocessors and Peripherals, CMOS VLSI Design, Advanced Networking Technologies, Mobile Communication

SKILLS

- **Expertise:** Microarchitecture, caches, coherence protocols, memory controllers, interconnects, address translation, SIMD processors, and GPUs
- **Programming Languages**: C/C++, Python, Bash Shell Scripting, System C/C++ (TLM 2.0), RTL Verilog, MIPS Assembly Language, CUDA C/C++, C++ Multi-thread, VHDL, POSIX (Pthread), MPI, Embedded C
- **Software Practices**: Test Driven Development, Agile Software Development, SCRUM BAN, Clean Code
- **Software Tools/Libraries/Frameworks**: GEM5, ChampSim, CATCI, McPAT, gdb, JIRA, synopsys Design Compiler, OmniGraffle, MATLAB, Asitic, ADS, MARS, ns3, Mininet, Wireshark Packet Analyze, GNU Debugger, MPLAB, OrCAD Capture, LTSPICE, LabVIEW, Xilinx ISE, AutoCAD
- **Protocols**: UFI, CMI, MAC, WiFi IEEE 802.11, UDP, TCP/IP, SPI, UART, I2C, Bluetooth, ZigBee IEEE 802.15.4, USB, CAN
- **Operation System Platform**: Linux/Unix, macOS, Windows

WORK EXPERIENCE

Advanced computing lab, Samsung Semiconductor — San Jose, CA
SoC Performance Architect — **May 2023 – Present**

- Developed performance model including Arm CPUs, caches, coherent interconnect, memory controller, and prefetchers. Implemented cycle accurate cache, interconnect, and memory controller unloaded/loaded latency based on cache pipelines delivered by RTL team.
- Performed correlation study of cache latency/bandwidth, Geekbench5/6 scores, with a pre-silicon emulator to improve the accuracy of a performance model.
- Conducted cache hierarchy size sensitivity, playing a crucial role in deciding the next-generation memory subsystem.
- Developed memory controller prefetcher based on CPU sidebands which improved IPC by 2-4% for memory intensive SpecInt'17 for next generation Exynos SoC.
- Developed prefetcher destination that exploits memory level parallelism based on outstanding transaction queue occupancy.

Power and Performance, Intel — Hillsboro, OR
Client SoC Performance Architect Intern — **Nov 2021 – June 2022**

- Developed a performance model of NoC bridge for protocol translation for Client SoC to enable CPU and GPU integration.
- Analyzed different address map options for GPU workloads and recommended the best address map based on performance and bandwidth utilization.
- Analyzed fabric sensitivity study of GPU micros and propose NoC fixes on I2C bridge.

Electrical and Computer Engineering Department, Lehigh University — Bethlehem, PA
Research Assistant, Computer Architecture Laboratory — **Jul 2018 – May 2023**

- Designed new Scratchpad Memory for general purpose CPU to increase security against cache-based side channel attacks, accelerate log lookups in persistent transactions, and increase performance of embedded benchmarks using shadow address space. (HiPEAC'21)
- Designed heterogeneous CPU-IMAC architecture to realize energy and performance improvements for CNN inference in mobile devices. (ISVLSI'21)

- Designed HW/SW co-design architecture to compute SpGEMM efficiently without requiring complex interconnection networks along with fast packing algorithm, SorPack, to convert a sparse matrix into a dense matrix that increases PE utilization. (HPCA'23)

Teaching Assistant, Digital Systems Laboratory **Jan 2020 – May 2020**

- Guided labs on using Verilog to implement hierarchical design, asynchronous circuits, vector-vector multiplication, memory modeling, and mapping algorithm (sorting and LCS) on Nexys-4 DDR FPGA using Xilinx Vivado HLx.

Electrical and Computer Engineering Department, University of Rochester Rochester, NY

Research Assistant, Advanced Computer Architecture Laboratory **Jan 2017– May 2018**

- Designed a new Cache Replacement Policy to eliminate dead blocks and to achieve low miss rate at Last Level Cache, tested on SPEC 2006 Benchmarks.
- Worked with Hardware Prefetchers to get better speed up compared to present ones, on GEM5 and tested SPEC 2006 Benchmarks with it using Python and C++ language.

Teaching Assistant, Computer Organization **Jan 2017– May 2017**

- Oversaw labs on using Verilog to design a register file, an ALU, a multiplier, a divider, and a single-cycle MIPS processor.
- Supplemented student learning by holding office hours and brainstorming sessions with student groups.

Teaching Assistant, Advanced Computer Architecture **Aug 2017 – Dec 2017**

- Guided labs on using Verilog to design a two-issue out-of-order processor, which can run real applications in Verilator, an open-source Verilog simulator.
- Designed benchmarks for testing and grading the labs.

Teaching Assistant, Probability for Engineers **Aug 2017 – Dec 2017**

- Assisted students in implementing various Probabilistic models using MATLAB Programming during labs.
- Graded exams, homework and cleared doubts of topics taught during course.

Research Assistant, Wireless Communication and Networking Group **Jan 2017 – Aug 2017**

- Worked on D2D (device-to-device) communication Physical layer, Mac layer and Network layer of Wi Fi-Direct in Python where the aim is to create Wifi-Direct environment and to implement Multi WLAN D2D Communication with Adaptive Routing Protocol between different WLAN.

Teaching Assistant, Circuits & Signals **Jan 2017 – May 2017**

- Assisted students in implementing various Discrete Signal Circuit models using MATLAB Programming during labs, grade labs, grade homework, clear doubts of topics taught, and grade exams for a class of 75 students.

Mumbai University Mumbai, India

Writer/Editor for College Magazine **Sept 2012 - May 2013**

- Edited articles and posts written by students for annual year magazine.
- Worked with alumni as well as current students to ensure everyone is correctly represented in magazine.

Marketing Team Coordinator **Sept 2012 – May 2013**

- Managed events, ranging from public relations, meeting with principals of other universities to attract students for our events, and held weekly meetings with event coordinators to gauge event status and formulate better marketing strategies.
- Created posters for events held using PowerPoint which attracted 20% more participants than previous year.

PUBLICATIONS

- **AMC: Access to Miss Correlation Prefetcher for Evolving Graph Analytics, Abhishek Singh**, Christian Schulte, Xiaochen Guo. (Under Review)
- **HIRAC: A Hierarchical Accelerator with Sorting-based Packing for SpGEMMs in DNN Applications,** Hesam Shabani, **Abhishek Singh,** Bishoy Youhana, Xiaochen Guo, 2023 IEEE International Symposium on High Performance Computer Architecture. (HPCA'23)
- **SPX64: A Scratchpad Memory for General-Purpose Microprocessors**, **Abhishek Singh**, Shail Dave, PanteA Zardoshti, Robert Brotzman, Chao Zhang, Xiaochen Guo, Aviral Shrivastava, Gang Tan, and Michael Spear, Transactions on Architecture and Code Optimization (TACO), Dec. 2020. (Presented at HiPEAC'21)
- **An In-Memory Analog Computing Co-Processor for Energy-Efficient CNN Inference on Mobile Devices**, M. Elbtity, **A. Singh**, B. Reidy, X. Guo, and R. Zand, ISVLSI'21 BEST PAPER AWARD.
- **Intelligent and Interactive Chess playing Robotic Arm Against Humans Project,** 2016 Tata Consultancy Services. (TCS)
- **Touch Screen Based Automated Medical Vending Machine,** National Institute of Technology (NIT), Nagpur and Dwarkadas J Sanghvi College of Engineering, India in 2015.

POSTERS

- **An In-Memory Analog Computing Co-Processor for Deep Learning at the Edge**, **Abhishek Singh**, Brendan C. Reidy, Xiaochen Guo, and Ramtin Zand, IBM IEEE CAS/EDS AI Compute Symposium (AICS), Oct. 2020.

RELATED ACADEMIC PROJECTS

- **HIRAC: A Hierarchical Accelerator with Sorting-based Packing for SpGEMMs in DNN Applications, C++:** Designed a novel sorting-based packing algorithm, SorPack, and a tile-based hierarchical SpGEMM accelerator. SorPack increases PE utilization by keeping the partial sums that need to be added together close to each other. The accelerator is a scalable system that maximizes the parallelism of the PEs. It achieves an average of 3.2× speedup on a single layer of DNN as compared to the state-of-the-art sparse DNN accelerator SIGMA with 9.5% area reduction and a 32% power reduction. An end-to-end evaluation on a DNN model shows an 8.2× runtime reduction over the TPU.
- **SPX64: A Scratchpad Memory for General-purpose Microprocessors, gem5, C++:** Designed a new scratchpad memory a **hardware-software** co-design for general purpose CPU which brings many benefits, including increased security and improves performance, especially for workloads with high locality or that interact with nonvolatile memory. It provides security against **Cache-based Side Channels and Spectre (Variant 1)** with 2× more performance as compared to speculation safe baseline. It provides about 1.12× speedup over unsafe baseline as compared to state-of-the-art MounTrap that performs same as unsafe baseline. For persistent transactions, SPX64 is used to **accelerate log lookups** by leveraging the virtual-addressing and set associative feature of SD$ by 1.16× compared to conventional processor. SPX64 accelerates embedded benchmarks by 1.10× using SD$ as **shadow address space** as compared to conventional processor**.**
- **An In-Memory Analog Computing Co-Processor for Energy-Efficient CNN Inference on Mobile Devices, ChampSim C++:** Designed an in-memory analog computing (IMAC) architecture realizing both synaptic behavior and activation functions within non-volatile memory arrays. Designed **heterogeneous** mixed-precision and mixed signal **CPU-IMAC** architecture to realize energy and performance improvements for CNN inference in mobile devices. It exhibits 6.5% and 10% energy savings for CPU-IMAC based realizations of LeNet and VGG CNN models, for MNIST and CIFAR-10 pattern recognition tasks, respectively.
- **Optimization and Parallel Architecture, C++:** Designed a new **Cache Replacement Policy** in gem5 framework which has on average 5.7% more hit rate on SPEC2006 benchmarks at Last Level Cache over LRU policy. Implemented **Next-Line** Hardware Prefetcher and **Best-Offset** Hardware Prefetcher in **gem5 simulation framework** which resulted an increase of 5.28% speedup for **Next-Line** Hardware Prefetcher over no prefetch and increased speedup of 10.995% for **Best-Offset** Hardware Prefetcher over AMPM Prefetcher on SPEC2006 benchmarks. Implemented **T2** Hardware Prefetcher in gem5 which turns out to be 84% accurate as of to 34-64% with above mentioned Hardware Prefetchers.
- **Principal Component Analysis on GPU, Cuda C++:** team of 2, implemented PCA algorithm for dimensional reduction of large input data for machine learning application which gain speedup of 4× compared to parallel CPU version of python PCA.
- **Mobile Ad-Hoc Network Optimization Through Interaction Visualization and Evaluation, Python:** implemented Multi-Hop, Multi-WLAN and Multi-Interface **WiFi-Direct** Network Topology on Mininet emulator using **Optimized Link State Routing Protocol** as Routing Mechanism between different Wireless LANs. Experimental results demonstrate superiority of techniques when compared to **WiFi-Mesh** that exploit device ability to maintain simultaneous physical connections to multiple groups, enabling multi-hop ad hoc networks with low overhead which is further beneficial for **Device-to-Device** Communication.
- **Exploiting Multithreading, C:** implemented **Pthread** and **MPI** technique to reduce execution time respecting various Coherence and Consistency models to try to achieve Maximum **Amdahl's Law** Limit. Outcome of using these Techniques was to imply own logic and to exploit Simultaneous Multithreading Technique in **C**.
- **Out of Order MIPS Processor, RTL Verilog**: team of 2, implemented Out of Order (**OoO**) MIPS processor with following modules files such as decodequeue1.v which is 8-entry, FIFO, loadstorequeue1.v which is 8-entry, FIFO, renamequeue1.v that is 8-entry, FIFO, Rrat.v consists of 64 entries, Frat.v which is 64 entries, RetireCommit.v includes 64 entries, Issue Queue that is 16-entry to achieve Practical speed up to 10.14% as compared to In-Ordered MIPS processor which passed ASM and CPP tests successfully.
- **Long Range Wide Area Network (LoRaWAN), an overview and its importance in the world of IoT, C++:** team of 2**,** focuses on new connectivity Protocol known as LoRa which is gaining importance in world of IoT for its capability to communicate over a long range while exploiting low power advantages of conventional Low Power Wide Area Network protocol (LPWAN).
- **Designing of Caches in Five Stage Pipelined Processor, RTL Verilog:** team of 2, designed 4 MB 8-Way Associative Cache which is L2 Cache shared by 32 KB Direct Mapped L1 Instruction Cache and 2-Way Associative L1 Data Cache, Block Size is 32 Byte for a Pipelined Processor. Outcome was practically implemented cache and observe speed up of 14.14% when there is only Main Memory no cache in 5-stage Pipelined Processor. Successfully passed CPP and ASM tests.
- **Five Stage Pipelined Processor, RTL Verilog:** team of 2 designed 5-stage pipeline microprocessor with control and data hazard detection and forwarding units. Design is implemented in an incremental fashion with different components being designed at different stages. Tested processor on number of test programs successfully.

- **Intelligent and Interactive Chess playing Robotic arm against humans, Embedded C:** team of 3, robot uses Cartesian coordinates to intelligently predict moves of human opponents in game of Chess. Programming in **ARM M3** processor, **FRDM-KL25Z** and Camera module **OV7670**.
- **Data Collection and Analysis**: team of 12, collected data and generated surveys to plan necessary changes required in conditions of Dharavi, largest slum in Asia**,** into a tourist spot to collect funds and increase GDP. Future scope of project is to help create small scale industries in poor and developing areas.
- **Mobile Charging Device Using Coin Module System, C:** team of 3, designed charging device which operates on coins, giving a certain amount of time for charging per input of coins. Arduino Board was used.
- **Transistor Tester using Discrete Components:** team of 3, developed using Discrete electronics to check operability of BJT transistors

AWARDS

- **ISVLSI'21, BEST PAPER AWARD** for An In-Memory Analog Computing Co-Processor for Energy-Efficient CNN Inference on Mobile Devices, M. Elbtity**, A. Singh**, B. Reidy, X. Guo, and R. Zand.
- 2021 Rossin Professional Development Program Award by Lehigh University.
- ISCA 2021 Registration Award.
- ISCA 2019 Travel grant from IEEE TCCA funds.
- **TCS'16 Best Student Project Award** for Intelligent and Interactive Chess playing Robotic Arm Against Humans**,** by TATA Consultancy Services.