

Ayush Kumar Sinha

☎ +91-(81972)-48927 — ✉ ayushkrsinha.work@gmail.com — 🔗 linkedin.com/in/ayush-s-626024251 — 🏠 Hyderabad, Telangana, India

Professional Summary

- Senior Deep Learning Compiler Developer with over 4 years of experience specializing in hardware accelerators
- Adept at designing and optimizing compilers for CPUs, GPUs, FPGAs, and TPUs using LLVM, MLIR, and TVM frameworks
- Proficient in C++ and Python with a strong understanding of advanced hardware architectures and machine learning kernels
- Experienced in building system software for domain-specific architectures
- Expertise in C++ parallel programming, concurrent programming, socket programming, and IPC
- Hands-on experience with modern C++, build systems, HPC libraries, testing frameworks, and debugging tools
- In-depth understanding of Linux OS and computer architecture
- Converted internship into a full-time position 3 months prior to its conclusion
- Focused and goal-driven with strong work ethics and a commitment to delivering quality work
- Excellent communicator with strong analytical, problem-solving, and organizational skills

Work Experience

AMD

Feb 2023 - Present

Senior System Software Engineer - AI

- Optimized the performance of Gen AI and CNNs on cutting edge hardware accelerators.
- Led the debugging, optimizing, profiling, testing and enhancement of compiler functionalities, resolving critical bugs while integrating sophisticated features such as code enhancement, new operator support, and optimization passes. These contributions markedly elevated the compiler's performance and reliability.
- Engineered specialized hardware-specific optimizations encompassing graph transformations, data types and layouts transformation, and memory access patterns. These optimizations were meticulously tailored to fully leverage the computational capabilities of designated hardware architectures, thereby optimizing overall computational efficiency.
- Conceptualized and executed the development of utility tools within the compiler to visualize aspects such as tensor memory allocations, tiling and the transformed intermediate representations as graphs. This graphical user interface tool provided pivotal insights into memory utilization and graph transformations, thereby enhancing the effectiveness of optimizations and simplifying troubleshooting processes.
- Developed an intricate custom logging framework within the compiler to systematically capture extensive runtime data and system states. This tool significantly streamlined the debugging process by enhancing the visibility of internal compiler operations, thus facilitating more efficient profiling and expedited resolution of performance bottlenecks.
- Accelerated team onboarding by documenting compiler frontend parsing and tiling engine. This effort significantly enhanced our team's understanding of the codebase, streamlined onboarding, and expedited development processes.

🏆 Recognised with the AMD Spotlight Award

AlphaICs

Oct 2021 - Jan 2023

Senior Deep Learning Compiler Engineer


- Spearheaded the enhancement of deep learning compiler strategies by deploying sophisticated techniques such as node fusion, node fission, auto-tiling, graph partitioning, and dead node elimination, integrated with dynamic shape inference capabilities. These optimizations collectively facilitated a 22% reduction in the computational overhead for inference tasks across various classification and detection neural networks.
- Orchestrated the augmentation of the existing DL compiler's framework compatibility, successfully integrating seven additional deep learning frameworks. This was achieved by harnessing the power of TVM's unified Relay IR, optimizing the compiler's ability to translate high-level computational graphs into optimized machine-level code tailored for a proprietary AI accelerator.
- Directed the design and implementation of a multithreaded assembler for superscalar processors, which streamlined the code compilation process, yielding a 4% average reduction in compilation times across projects.
- Engineered a multithreaded disassembler, GUI debugger and tiling visualizer for an AI accelerator. This initiative significantly enhanced the efficiency of developing and troubleshooting neural network kernel libraries, thus accelerating deployment cycles.
- Developed a simulator for AI accelerator to support robust CI/CD verification processes for software repositories, ensuring higher code quality and reliability.

Machine Learning Kernel Development

- Optimized Kernels for Neural Network Operations on Specialized Hardware - Meticulously tuned a suite of kernels for key neural network operations — including convolution, pooling, up/down-sampling, activation, reduction, and mathematical algebra functions—to achieve optimal performance on specialized hardware. Utilized advanced

techniques such as instruction pipelining, enhanced instruction selection, and minimized memory transactions to maximize efficiency.

- Reduced ResNet-50 Execution Time by 96 Milliseconds through Precision Engineering - Achieved a significant 96-millisecond reduction in the execution time of the ResNet-50 model by applying precision engineering on our custom hardware platform. Benchmarked performance gains validated the effectiveness of advanced optimization techniques, leading to substantially faster model inference.

 Recognised with the Employee of the Month Award

AlphaICs

Jun 2020 - Sep 2021

Deep Learning Compiler Engineer

- Developed efficient utility subroutines for digital image processing algorithms, enabling the Development Team to integrate advanced functionalities seamlessly without delving into underlying complexities.
- Conducted in-depth critical analyses of hardware and software implementations of activation functions and matrix multiplication, delivering comprehensive viability reports that influenced strategic decisions on microarchitecture optimization.
- Authored and maintained comprehensive documentation for libraries, Instruction Set Architecture, and microarchitecture, streamlining knowledge transfer and accelerating onboarding for colleagues and mentees.

Education

Jadavpur University - Bachelor of Technology

2017-2021

Grade: First Class

Undergraduate Projects:

- Digital Image Stitching for Creating High-Resolution Panoramas.
- Steganography for Secure Transmission of Multimedia Data.

Undergraduate Seminar:

- The Algorithms Behind Adobe Photoshop's Blending Modes and Filters.

Activities and Societies:

- Placement, Entrepreneurship, & Incubation Cells
- Code & Robotics Club
- Lead Organizer, Annual Cultural & Tech Fest
- Workshop Presenter: Led workshops focused on game/web/app development and ethical hacking.