# Shlok Limbhare

*AI Engineer — GPU Systems — CUDA Optimization*
Shlokvfx2003@gmail.com — +91-7777-885-887
Dream Castle, Nashik, MH — GitHub: ShlokVFX — LinkedIn: ShlokLimbhare

*Keywords: GPU Kernel Programming, CUDA, FP8, MI300X, Matrix Multiplication, Memory Optimization, High-Performance Computing, MoE, MLA Decode, Triton, cuBLAS*

## Projects

**100 Days of CUDA: Optimized GPU Kernel Programming** GitHub

- Documented a structured 100-day learning journey in CUDA, covering core concepts, memory hierarchy, and kernel optimizations.
- Developed and profiled high-performance CUDA kernels with Streams, Shared Memory Tiling, Unified Memory, and Loop Unrolling to enhance execution efficiency.
- Implemented Tensor Core acceleration with WMMA, explored fused operations in Triton, and optimized cuBLAS-based linear algebra routines.
- Designed custom GPU-accelerated implementations for activation functions (Softmax, ReLU) and performance-optimized numerical kernels.
- Leveraged CUDA Graphs and `torch.compile` for reduced computational overhead in deep learning pipelines.
- Achieved **3x–5x speed improvements** in various CUDA-based workloads, demonstrating expertise in multi-GPU computing and performance optimization.

**AMD MI300X GPU Kernels: FP8 MatMul, MoE, MLA Decode** GitHub
*Contributor — HIP, FP8 Kernels, MoE Inference, Popcorn Eval*

- Developed HIP-based double-buffered matrix multiplication kernels supporting FP8 blockwise computation using MFMA intrinsics on AMD MI300X.
- Built optimized inference modules for Mixture-of-Experts (MoE) and MLA (Multi-head Latent Attention) decoding targeting large model inference.
- Benchmarked performance using Popcorn's eval harness on MI300X hardware to validate throughput, latency, and scaling behavior.
- Tuned shared memory usage, vectorized tile access, and kernel launch configurations to match AMD architecture design.
- Demonstrated expert-level understanding of ROCm, FP8 formats, shared memory pipelining, and GPU-specific compiler flags for HPC workloads.

## Technical Skills

**GPU Programming:** CUDA, HIP, Triton, OpenCL, Nsight, WMMA, MFMA
**Frameworks:** PyTorch, TensorFlow, Triton, TorchScript
**Performance Tools:** cuBLAS, cuDNN, CUDA Graphs, Popcorn, Perfetto
**Languages:** Python, C++, Docker, Git, CMake
**Model Types:** MoE, LLM Inference, MLA Decode, Transformer Ops
**Platforms:** AMD ROCm (MI300X), NVIDIA RTX, Linux HPC Clusters

## Technical Writing & Talks

**Optimizing LLM Inference on AMD MI300X: FP8 and MoE Kernels** (Medium Blog, 2025)
Covers FP8 matmul, shared memory tuning, and MoE-based large language model inference on MI300X.

## Professional Experience

**Technical Artist** *DNEG, Mumbai, MH*
2021 – 2024

- Developed VFX production tools using OpenCL, C++, and VEX shader programming.
- Automated repetitive tasks via scripting, reducing manual workload by **40%**.
- Issued weekly reports optimizing software performance and rendering pipelines.

## Education

**B.Sc. Computer Science** *BITS Pilani, Pilani, RJ*
2023 – 2027

- Specialized in Artificial Intelligence, Machine Learning, and High-Performance Computing.
- Completed coursework in Data Structures, Algorithms, Operating Systems, and Distributed Systems.
- Led academic projects involving GPU-accelerated deep learning pipelines and compiler-aware optimizations.

## About Me

GPU Kernel Engineer focused on accelerating inference workloads and pushing the limits of hardware efficiency. I build optimized compute kernels for CUDA and AMD ROCm platforms, with deep interest in low-precision arithmetic, HPC inference, and end-to-end pipeline profiling for large models.