Darshan Baslani

+91 77779 33517 | dcbaslani@gmail.com |
Portfolio | LinkedIn | Github

Summary

Intensely curious and passionate ML Systems Engineer driven by a first-principles approach to AI inference engine design and low-level GPU hardware architectures. Expert at tracking down framework bottlenecks via profiling and writing highly optimized, memory-coalesced kernels to saturate hardware compute boundaries. Proven track record of accelerating linear attention models and autoregressive loops, leveraging Triton, Tile_lang, and CuTe abstractions to deliver near-theoretical peak throughput on state-of-the-art accelerators.

Technical Skills

Low-Level & Systems: C, C++, CUDA, Python, Go, Linux Namespaces, cgroups
ML Inference & Compilers: PyTorch, CuTe, Triton, Tile_lang, Custom Autograd Engines, vLLM Architecture
Performance Engineering: PyTorch Profiler, Nsight Compute, Kernel Fusion, Shared Memory Swizzling, TMA, WGMMA

Inference & Systems Projects

Qwen3.5 High-Throughput Inference Engine | *Triton, Tile_lang, PyTorch Profiler* Blog | Github

- Achieved a **5x end-to-end token generation speedup** (from 16 to **83 tokens/sec**) optimizing Qwen 3.5 on an NVIDIA Blackwell B200 instance by resolving memory-bandwidth saturation and framework graph bottlenecks.
- Re-architected the transformer block boundaries into a **vLLM-style continuous residual stream** by passing structured activation-residual tuples, completely eliminating global memory (HBM) roundtrips between decoder layers.
- Designed and implemented a custom **Fused Zero-Centered RMSNorm** in **Triton** yielding a **5.7x reduction** in kernel latency, and ported the Gated Delta Net chunked recurrence rule to **Tile_lang** to explicitly optimize asynchronous TMA loads and swizzled SRAM tiles.

Blackwell B200 GEMM Optimization & CuTe Architecture | *CUDA, CuTe* Leaderboard | Github

- Ranked **#40 globally** in the GPU Mode group competition by directing a custom AI agent to successfully generate a high-performance Group GEMM kernel leveraging the experimental **NVFP4** datatype on the Blackwell (B200) architecture.
- Developed a custom SGEMM kernel that achieved **110.6% of NVIDIA cuBLAS performance** on local consumer hardware (GTX 1650) by aggressively optimizing shared memory carveout, 2D register tiling, and global memory coalescing.

Gen AI Inference Optimization (Denoising & Autoregressive) | *PyTorch, CUDA* Blog | Github

- Built **Stable Diffusion 1.5** from first principles (UNet, VAE, CLIP) to profile denoising bottlenecks; reduced inference latency by **85%** (33s to 4.7s) via **Flash Attention-2**, and FP16 quantization.
- Engineered **Simple-LLM**, a from-scratch GPT-style autoregressive model, to experiment with efficient decoding strategies and memory-bound inference constraints.

Low-Level ML Frameworks & Compilers | *CUDA, C++* Github | Blog

- Developed **Vibegrad**, a custom autograd library built from scratch, demonstrating a deep mechanical understanding of computational graphs and automatic differentiation used in modern ML compilers.
- Engineered an **Online Softmax** CUDA kernel that outperforms native PyTorch implementations by utilizing warp-level reduction primitives, resolving shared memory bank conflicts via **Nsight Compute** profiling.

Systems Engineering | *Go* Github

- Programmed a custom **Container Runtime** in Go to master OCI standards, utilizing Linux `namespaces` for process isolation and `cgroups`, proving deep knowledge of embedded and OS-level virtualization constraints.

Technical Writing & Publications

dcbaslani.xyz/blog

- **Performance Optimization Worklogs:** Authored a comprehensive deep dive on accelerating Qwen 3.5 linear attention mechanisms on Blackwell architectures using Triton and Tile_lang; documented first-principles optimization worklogs for Stable Diffusion 1.5 and fast CUDA Online Softmax implementations.
- **Deep-Dive CUDA & CuTe Series:** Authored an extensive, 9-part technical series on modern GPU programming using the CuTe library. Topics range from foundational memory layouts to building a custom **CuTe DSL**, exploring advanced features like Warpgroup MMA (WGMMA), and the TMA (Tensor Memory Accelerator).
- **Computer Science & Math Foundations:** Published foundational articles detailing the building blocks of Propositional Logic and constructing Linear Regression entirely from scratch.

Experience

AI Engineer Intern Feb 2025 – April 2025
Edutor App (1M+ Users) *Palanpur, Gujarat*

- Architected a real-time voice-conversation agent, optimizing ASR/TTS runtime pipelines for low-latency, real-world vernacular interactions.
- Led the deployment of models to **Serverless Microservices** on GCP Cloud Run, adapting infrastructure to variable networking and compute constraints.
- Built asynchronous request handling tooling and caching layers, reducing production inference latency by 40%.

ML Engineer Intern May 2025 – July 2025
Machine Learning Studies *Remote*

- Designed a high-throughput Retrieval-Augmented Generation (RAG) pipeline; benchmarked embedding models and optimized memory-bound indexing parameters for reliable sub-second retrieval.

Education

Dayananda Sagar University Present
Master of Science (M.Sc.) in Data Science *Bangalore, India*

B.K. Mehta College June 2022 – Mar 2025
Bachelor of Computer Applications *Palanpur, Gujarat*