MARCO HARUNI

ML Systems & LLM Inference Engineer

Dar es Salaam, Tanzania | Open to remote work and international relocation
mtaniharuni95@gmail.com | Website | GitHub | LinkedIn | Hugging Face | X

TECHNICAL SKILLS

Languages: Python, CUDA C++, C++, Rust, Bash
GPU & ML: PyTorch, JAX, Flax NNX, Optax, Triton, Pallas, CuTe DSL, vLLM, Hugging Face Transformers
Inference Systems: prefill/decode, paged KV cache, continuous batching, GQA/MQA, online softmax, sampling, FP8 KV cache, SSE streaming
Engineering: Linux, Docker, Git/GitHub Actions, pytest, uv, Modal, profiling, benchmarking, numerical correctness

OPEN-SOURCE & INDEPENDENT ENGINEERING

- **Active portfolio:** maintain ongoing public ML systems repositories with regular improvements to kernels, correctness tests, benchmarks, and documentation.
- **Ecosystem work:** implemented a real out-of-tree vLLM custom operator and work directly with vLLM, SGLang, TensorRT-LLM, and FlashInfer codebases and interfaces.

SELECTED OPEN-SOURCE PROJECTS

ForgeEngine | 2026-Present

- Build and maintain a readable single-GPU Qwen3 4B inference engine with staged SafeTensors loading, explicit prefill/decode, paged KV caching, token-budgeted continuous batching, sampling, incremental detokenization, and a streaming chat API.
- Validated on an NVIDIA L4 with 93 Python tests, 16 parameter checks, and 3 Rust tests; measured 15.62 generated tokens/s at concurrency 4, 8.88 GB peak CUDA allocation, zero failed requests, and zero leaked KV blocks.
- Implemented Triton RMSNorm and CUDA paged-GQA paths with PyTorch fallbacks and deterministic cancellation cleanup.

LLM Serving Kernels | 2026-Present

- Implement eight inference-kernel projects in Triton, CUDA C++, CuTe DSL, and vLLM covering RMSNorm, SwiGLU, GEMM, paged decode attention, FP8 KV cache, top-k sampling, MoE, and an out-of-tree custom operator.
- Passed 51 L4 GPU correctness tests; measured 1.71x RMSNorm and 5.12x paged-decode speedups, plus 49.9% lower KV storage at 0.00110 mean absolute decode error.
- Published complete baselines and limitations, including a correct vLLM operator that remained 9.7% slower than stock vLLM.

Pallas MoE Kernels | 2026-Present

- Develop a fused approximate-GELU routed-MoE forward kernel in JAX Pallas that avoids materializing the full hidden tensor and accumulates tiled projections in FP32.
- On an NVIDIA L4, achieved 2.05x expert-core and 1.84x end-to-end speedups with approximately 81% lower XLA peak memory; passed 36 local and 9 GPU tests across FP16/BF16 edge cases and gradients.

JAX Addition Transformer / Hugging Face | 2026-Present

- Built an exact 10,000,000-parameter decoder-only transformer from scratch with JAX/Flax NNX; trained on a Colab T4 in 6.89 minutes after compilation and achieved 1,000,000/1,000,000 greedy autoregressive exact matches with zero invalid generations.
- Completed a 32-run dense-versus-MoE scaling study and a 17.94M stored-parameter top-2 ragged-dot MoE baseline.

ADDITIONAL OPEN-SOURCE WORK

- **Forge LLM** - configurable 88.6M-parameter PyTorch pretraining reference with GQA, RoPE, QK-Norm, fused QKVO, token-budgeted training, checkpoint/resume, metrics, and generation.
- **Forge Tokenizer & Forge Position** - byte-level BPE, Unicode normalization, positional encodings, RoPE variants, ALiBi, context scaling, and KV-cache position handling.

EDUCATION

Muhimbili University of Health and Allied Sciences (MUHAS) | Dar es Salaam, Tanzania
Bachelor of Science in Physiotherapy, Expected 2027
Independent ML Systems Study, Ongoing - deep, project-based specialization in transformers, model training and inference, GPU programming, profiling, and open-source serving systems.

Public repositories link directly to source code, tests, benchmark evidence, and documented limitations.