

Nicholas Namusanga

GPU KERNEL ENGINEER · CUDA KERNEL DEVELOPMENT & OPTIMIZATION · PTX/SASS LEVEL TUNING · 10 YRS EXPERIENCE

hey@namusanga.com | namusanga.com | linkedin.com/in/nichwithn | <https://github.com/nnamu-cl> | Kampala, UG · remote, UK hours

I build and optimise CUDA kernels and GPU pipelines for AI and HPC workloads: kernel and compute-pipeline development, memory optimisation and profiling for peak throughput. 10+ years in low-level C++ on compute and memory bound workloads across AI compute, real-time video, simulation and numerical methods.

EXPERIENCE

My career started with creating GPU resident systems for realtime simulators then progressed to CUDA compute for modern HPC and AI workloads.

Senior HPC Performance Engineer / Jugo †

Mar 2026 - Sept 2026

Contract · United Kingdom, Remote

- **Built and modified** custom CUDA kernels for live video processing on Jugo's immersive meetings platform, which renders real-time 3D meeting environments in the browser on an NVIDIA and Unreal Engine stack.
- **Work across the NVIDIA DeepStream** video stack, taking a game-engine render pipeline and adapting it to live AI video compute and real-time, low-latency streaming under a hard per-frame budget.
- **Profile and optimize** the compute and memory bound stages of the pipeline for throughput at scale, from one-to-one calls up to town-hall audiences, using Nsight to drive occupancy, launch configuration and memory-coalescing changes.

GPU Compute Engineer / PSI Engine (CUDA physics & simulation) †

2024 - 2026

C++ · CUDA · Vulkan

- **Built** a node-based physics, compute and simulation engine from scratch in C++. Physics pipekines are represented as a graph of nodes; the engine evaluates the DAG, strategically runs numerical work on a CUDA pipeline and returns results over device memory to a Vulkan render engine. The compute path utilizes high stream concurrency and custom kernel implementations to exploit the workloads at and maximize throughput.
- **Designed and wrote** custom sparse linear-algebra CUDA kernels: notably, a delta-input sparse matrix multiply that maintains a dynamic patch propagating over t on the GPU to achieve performance boosts over static column selection and cuSPARSE stateless compute on specific dirty-set propagations typical in power-flow, circuit and implicit-FEM solvers.
- **Profiled** the hot kernels with Nsight Systems and Nsight Compute and tuned shared-memory tiling, occupancy and memory coalescing for sustained bandwidth.

Lead Graphics Engineer / Union One

Aug 2018 - Aug 2024

Contract · United Kingdom, Remote

- **Delivered** the GPU-resident render pipeline for Union One, a tactical FPS: particle systems and related effects computed on the GPU rather than the CPU, with Vulkan driving rendering and the surrounding engine systems.
- **Owned** subsystems where frame budget and memory layout were the binding constraints, profiling and optimising against a fixed per-frame time budget.

TECHNICAL

LANGUAGES C++, CUDA, GLSL, Python

SKILLS CUDA kernel development, Vulkan, NVIDIA DeepStream, parallel algorithm design, memory & cache optimisation, Nsight profiling, system & application performance

CERTIFIED CUDA Programming, Johns Hopkins University