JASHOM TECHNOLOGIES PVT. LTD.

Precision GPU Engineering for High-Performance AI

We build the infrastructure that makes AI faster, leaner, and more powerful — from CUDA kernel-level optimisation to full-stack GPU pipeline delivery.

GPU OPTIMIZATION | CUDA DEVELOPMENT | HIRE GPU EXPERTS | RUST ENGINEERING | HPC SOLUTIONS

Company Brochure · 2025–2026

info@jashom.com
www.jashom.com

ABOUT US

Who We Are

Jashom Technologies is a specialist GPU engineering firm focused exclusively on high-performance computing, AI infrastructure, and parallel computing at scale. Founded by engineers who have worked on GPU architectures at the hardware and kernel level, we bring a depth of expertise that generalist IT providers cannot match.

We work with AI startups, enterprise data teams, research institutions, and cloud-native companies that demand measurable performance improvements — not just incremental gains. Every engagement is built around your specific workload, hardware constraints, and business goals.

Our Core Competencies

CAPABILITY 01

CUDA & GPU Kernel Engineering

We write, profile, and optimise CUDA kernels for maximum throughput. From memory access patterns to warp-level parallelism, we squeeze every FLOP from your hardware.

CAPABILITY 02

LLM Inference & AI Pipeline Optimisation

Quantisation, batching strategies, KV-cache tuning, tensor parallelism — we optimise the entire inference path for production LLM deployments at any scale.

CAPABILITY 03

Systems Programming (Rust & C++)

Our Rust and C++ engineers build low-latency, memory-safe systems for GPU orchestration, telemetry, and infrastructure control planes.

CAPABILITY 04

Staff Augmentation & Dedicated Teams

Need embedded GPU engineers? We supply dedicated CUDA and Rust developers who integrate directly with your team and your development workflow.

50+	12+	4+	3×
GPU PROJECTS DELIVERED	EXPERT GPU ENGINEERS	YEARS OF CUDA R&D	AVG. COST REDUCTION

OUR SERVICES

What We Do

SERVICE 01

NVIDIA GPU Optimization

End-to-end GPU performance engineering for AI, HPC, and data-intensive applications. We profile workloads, find bottlenecks, and implement kernel-level optimisations.

- Custom CUDA kernel profiling and rewrite
- Memory bandwidth and occupancy optimisation
- Multi-GPU and distributed inference tuning
- Power efficiency improvements

SERVICE 02

CUDA Development Services

Build scalable, production-ready GPU applications. Our CUDA developers design parallel algorithms for AI training, inference, simulation, and signal processing.

- Parallel algorithm design and implementation
- AI model acceleration (training + inference)
- Custom GPU libraries for scientific computing
- Integration with PyTorch, TensorRT, ONNX

SERVICE 03

Hire CUDA Developers

Dedicated CUDA programmers embedded in your team. Senior GPU engineers on monthly retainer or project basis with full IP transfer.

- Senior CUDA & GPU engineers on demand
- Monthly retainer or fixed-scope projects
- Full IP transfer and NDA-ready

SERVICE 04

Hire Rust Developers

Experienced Rust engineers for GPU orchestration, telemetry pipelines, and high-throughput systems. Memory-safe, fast, and production-proven.

- GPU orchestration and control plane systems
- High-throughput telemetry pipelines
- Memory-safe systems programming

PORTFOLIO

Real Results, Real Workloads

Every case study below represents a live production deployment. Numbers are measured post-optimisation under real workload conditions — not benchmarks.

OUTCOME · LLM INFERENCE

42%

THROUGHPUT IMPROVEMENT

LLM Inference Optimization on Constrained GPU Infrastructure

Deployed a 13B parameter model on a 12-node distributed cluster. Re-engineered the full inference path — CUDA kernels, dynamic quantisation, adaptive batching — delivering 42% higher throughput and 37% lower power, at S the projected cost.

OUTCOME · ORCHESTRATION

60%

IDLE GPU TIME REDUCED

GPU Workload Orchestration on Rocky Linux 9.7

Built a multi-queue GPU scheduling framework on Rocky Linux 9.7 that reduced idle GPU time by 60% and eliminated resource contention across concurrent AI training jobs.

OUTCOME · CLOUD GPU

3×

FASTER FINE-TUNING

Cloud GPU Fine-Tuning for Production LLM Deployment

Redesigned the fine-tuning pipeline for a production LLM on cloud GPU infrastructure. Mixed-precision training, gradient checkpointing, and dynamic batch scaling cut training time by 3× and cost by 55%.

OUTCOME · TELEMETRY

99.9%

HARDWARE VISIBILITY

Real-Time GPU Telemetry via Redfish BMC Integration

Integrated Redfish BMC telemetry with a custom Rust-based collector to deliver real-time GPU server health monitoring across a bare-metal cluster with sub-second latency.

TECHNOLOGIES WE WORK WITH

CUDA · TensorRT · PyTorch · ONNX · NCCL · Rust · C++17 · Rocky Linux · Kubernetes · Redfish BMC · H100 / A100 · NVLink

WHY JASHOM

The Jashom Difference

DEPTH

Kernel-Level Expertise

We work at the CUDA kernel and hardware level — not just framework wrappers. That depth is what produces 40%+ gains where others plateau.

FOCUS

100% GPU-Specialised

We don't do everything. We do GPU performance — exclusively. That focus means our engineers have solved the exact problem you're facing before.

DELIVERY

Measurable Outcomes

Every engagement starts with baselines and ends with verified benchmarks. We define success metrics upfront and hit them — or keep working until we do.

FLEXIBILITY

Your Model, Your Terms

Project-based, retainer, or embedded team — we work the way you work. Full IP transfer, NDA-ready, and timezone-aligned delivery.

Our Engagement Models

Fixed-Scope Project

Defined deliverables, timeline, and budget. Best for a specific optimisation task or feature build.

Monthly Retainer

Ongoing GPU engineering support — ideal for teams that ship AI features continuously and need expert capacity on call.

Dedicated Team

1–6 senior GPU engineers embedded in your team, working exclusively on your infrastructure and roadmap.

How It Works

- Step 1 — Discovery call: we understand your workload, hardware, and performance targets (30 min)
- Step 2 — Baseline audit: we profile your current system and identify the highest-impact optimisation opportunities
- Step 3 — Proposal: we deliver a scoped plan with timeline, team, and expected performance gains
- Step 4 — Execution: our engineers get to work, with weekly updates and live benchmark tracking
- Step 5 — Handoff: full documentation, code ownership, and optional ongoing support

GET IN TOUCH

Let's Talk Performance

Whether you have a specific GPU bottleneck, a scaling challenge, or just want to know what's possible — we're happy to talk. No sales pitch, just engineering.

EMAIL
info@jashom.com
Response within 24 hours

WEBSITE
www.jashom.com
Book a consultation online

PHONE
+91 90239 06363
Mon–Fri, 9am–6pm IST

OFFICE
Gandhinagar, Gujarat
SATYAM 1 414, Amba Business Park, Adalaj 382421

Powering High-Performance AI with Precision GPU Engineering

JASHOM TECHNOLOGIES PVT. LTD. · 2025–2026