

Arpan Pathak

Seattle, WA | Email: arpan.pathak47@gmail.com | Phone: +1 (206) 306-6059

LinkedIn: [linkedin.com/in/arpan-pathak-272341424](https://www.linkedin.com/in/arpan-pathak-272341424) | GitHub: github.com/arpanpathak

YouTube: [youtube.com/@ArpanPathak](https://www.youtube.com/@ArpanPathak)

Visa: Requires H1B Transfer sponsorship

SUMMARY

Senior Software Engineer with **8 years** building **AI inference serving**, **backend microservices**, and cloud-native infrastructure at **Amazon**, **Microsoft**, **Oracle**, and **Razorpay**. Strong in **Python**, **modern memory-safe C++**, **Rust**, and **CUDA**, with **BERT inference at 5,000 QPS** (ONNX, INT8) on **Docker/Kubernetes** with CI/CD. Passionate about **CV/DL for road safety** on **NVIDIA Jetson Orin** (TensorRT).

PERSONAL PROJECTS

Driving-CivicSense: Full-Stack Edge AI (Rust, CUDA), NVIDIA Jetson Orin Nano Super NX | [github](#)

- Collected **50,000 miles** of driving footage across all weather and terrain using vehicle camera sensors and **LiDAR**, generated simulation-driven synthetic data, and kept processing privacy-first and fully on-device on edge neural chips.
- Deployed **YOLOv8/v11 INT8** → **TensorRT** at **~12 ms** on **Jetson Orin Nano Super NX**, streamed sensor data to a **Raspberry Pi Zero 2 W UDP server**, and validated the system on **Raspberry Pi 5** and **Orange Pi Zero**.
- Building edge VLM serving with **vLLM** and **TensorRT-LLM** in Docker containers on the Jetson.
- Contributing GPU kernels and documentation fixes to **CUDA-Oxide (Rust-to-CUDA)**, wrapping unsafe intrinsics in explicit unsafe blocks, and pushing Jetson limits with idiomatic, memory-safe, fearlessly concurrent kernels.
- Authored a paper on **Intersection Blockage Prediction** ([paper](#)) and studied ISO 26262 safety standards. Books: **Seeing Machines · CUDA Kernels** ([book](#))

EXPERIENCE

Software Developer II, Microsoft, Redmond, WA

Jun 2025 – Aug 2026

- Built zero-trust sovereign-cloud networking with eBPF + Cilium enforcing **data residency** and SecNumCloud compliance, and wrote Kubebuilder operators that reconcile state across AKS clusters in France and Germany.
 - Built and maintained **data loss prevention (DLP)** and **data lifecycle management** services with a **C#/ .NET** backend.
 - Refactored a **legacy C++ codebase** with **modern C++** memory practices such as **smart pointers** and **RAII**, fixing **use-after-free** and **buffer-overflow** vulnerabilities to harden memory safety.
- TechStack:** C++, C#, .NET, Python, Kubernetes, Cilium, eBPF, Docker, CI/CD, Linux

Senior Software Engineer, Oracle Cloud Infrastructure, Seattle, WA

Oct 2024 – Mar 2025

- Owned the **Terraform provider** (developer SDK) for Autonomous Database, replacing legacy control-plane tooling and projecting **\$1.2M/month** in savings.
 - Built lock-free data structures that reduced contention and latency in high-concurrency read-write paths.
 - Integrated OCI Security Vault with automated key rotation across control-plane services.
- TechStack:** Java, Go, REST/gRPC, Terraform, OCI

Software Development Engineer II, Amazon, Seattle, WA & Hyderabad, India

Mar 2021 – Oct 2024

- Scaled low-latency **BERT-based inference** to **5,000 QPS** with ONNX Runtime and INT8 quantization, validated throughput and latency, and cut triage SLA from **7 days to 2 hours**.
 - Owned end-to-end training, serving, and latency tuning for an XGBoost and Deep Learning ensemble across **FreeVee, Twitch, MiniTV, Prime Video, and Amazon Retail**, lifting ad CTR **5%** and adding **\$5.4M/month** revenue.
 - Led a team of 3 building **programmatically guaranteed (PG) ad delivery** at **<50 ms** using real-time CDC in Kotlin with REST/gRPC APIs.
 - Designed a uniqueness-constraint indexing SDK with **two-phase commit**, **serializable isolation**, and **3K writes/s**.
 - Built the **1-petabyte** warehouse and lake with AWS Glue and Spark for financial reporting and pay computation.
 - Cut deal over-delivery by **~30%** and budget overspend by **25%** with real-time monitoring and automated controls.
- TechStack:** Python, Kotlin, PyTorch, XGBoost, ONNX Runtime, Spark, Kafka, DynamoDB, Redshift, AWS

SDE, Razorpay, Bengaluru, India

Aug 2020 – Feb 2021

- Built the P2P Unified Payments Gateway in Go with Protobuf/gRPC, sustaining **3M+ transactions/hour** at peak with **99.99% uptime** and **150ms P99** for customers like **Zomato, Swiggy, Zerodha, and Groww** on AWS and GCP.

Software Engineer, Mindfire Solutions, Bhubaneswar, India

Aug 2018 – Feb 2020

- Built Java/Spring full-stack tools and an AR product-search microservice handling **10K+ daily queries**.

SKILLS

Languages: Python, C++, Rust, Go, Java, Kotlin, C, Shell/Bash

AI/ML Inference: Deep Learning, Computer Vision, PyTorch, XGBoost, BERT, LLMs, ONNX Runtime, TensorRT (INT8), vLLM, SGLang, TensorRT-LLM, YOLOv8/v11, Edge Inference, Quantization, MLOps

GPU & Parallel Computing: CUDA, CUDA-Oxide (Rust-to-CUDA), SIMT, GPU Kernel Programming, NVIDIA Jetson Orin, HPC

Systems Programming & Performance: Lock-Free Data Structures, Concurrency, Memory Safety, Performance Engineering, Low-Latency Systems, Computer Architecture

Networking: TCP/IP, mTLS, HTTP/2/3, QUIC, eBPF, XDP, Cilium, Load Balancing

Distributed Systems & Streaming: Consensus, Replication, Sharding, Kafka, Spark, CDC, Microservices, gRPC, REST, Protobuf

Cloud & Infra: AWS (EC2, ECS, EKS, Lambda, Kinesis, EMR, SageMaker, CDK), Azure Kubernetes Service (AKS), Google Kubernetes Engine (GKE), Kubernetes, Docker, Helm, Terraform, Kubebuilder, Istio, CI/CD

Observability: Prometheus, Grafana, AWS CloudWatch

Data & Storage: DynamoDB, PostgreSQL, MySQL, Redis, Amazon Redshift, Amazon S3

EDUCATION

B.Tech, Computer Science & Engineering, RCC Institute of Information Technology, Kolkata | 2018