Aritro Roy

aritroroy1999@gmail.com • +1 (917) 615-7396 • Brooklyn, NY 11220 • linkedin • github

Education

New York University, NYC, New York, USA — May 2027
Master of Science, Computer Science **(3.889/4.0)**
Relevant Coursework: Advanced Operating Systems, Distributed Systems, Machine Learning, MLOps

Indian Institute of Technology, Dhanbad, India — Jun 2022
Bachelor of Technology, Electronics and Instrumentation Engineering

Technical Skills

Programming Languages: Java (SpringBoot), Python, TypeScript (Node.js)
Distributed Systems & Messaging: Kafka, Redis, BullMQ, SQS
Databases: MySQL, PostgreSQL, MongoDB
Cloud & Infra: AWS (Lambda@Edge, S3, CloudFront), Docker

Experience

Senior Backend Engineer, **Saathi WorldApp Pvt Ltd**, Gurugram, India — Jul 2024 – Aug 2025

- Designed an event-driven referral system processing 10K+ events/min for 1.3M+ users, supporting 100-level referral chains
- Owned production rollout and on-call reliability for event-driven services supporting 1.3M+ users
- Implemented versioned reward logic enabling zero-downtime config updates and live A/B testing
- Built a wallet service with atomic, idempotent transactions, achieving p99 API latency <150ms under peak load
- Designed and operated a fault-tolerant processing pipeline for CPU-intensive image and video workloads, executing 50K+ jobs/day in parallel across 4 microservices, with distributed tracing, exponential backoff retries, and DLQ-based isolation, achieving 99.9% job success rate
- Optimized image delivery using Lambda@Edge, CloudFront, and S3, cutting third-party costs by $1K+/month and improving load times with deep caching and multi-quality encoding
- **Tech Stack:** Node.js, MongoDB, Redis, Kafka, BullMQ, AWS, New Relic

Backend Engineer, **Intract Software Pvt Ltd**, Gurugram, India — Oct 2023 – Jul 2024

- Designed and deployed a real-time event indexing and alerting system over multiple L2 blockchains (Base, Scroll, zkSync, BNB), monitoring on-chain events to generate actionable market insights
- Architected a cross-chain stateful workflow integrating smart contracts, bridges, and swaps, improving user retention by 40% and increasing daily active wallets by 35%
- Led the design and implementation of a discovery and recommendation pipeline for digital assets, driving 5K+ daily transactions, 30% higher user engagement, and 25% revenue growth
- Launched and scaled an onboarding and engagement platform (10K+ DAU), implementing user fingerprinting and eligibility checks to reduce abuse by 95% and introducing incentive mechanisms that boosted engagement by 50%
- **Tech Stack:** Node.js, MongoDB, AWS, Solidity

Software Engineer, **BYJU'S (Think and Learn Pvt Ltd)**, Remote — Jun 2022 – Oct 2023

- Automated tutor allocation, eliminating a 2,000 man-hour manual process and scaling support for 10K+ tutors and 1M+ classrooms, reducing operational effort by 90%
- Designed and owned an Availability microservice to evaluate tutor eligibility and real-time availability, achieving 99.9% uptime and sub-100ms API response times
- Led cross-functional migration of business verticals to a shared horizontal platform, designing zero-downtime data migration strategies for 5M+ records with backfill and forward-fill, and deprecating legacy services while maintaining 99.99% data accuracy
- **Tech Stack:** Java Spring Boot, MySQL, MongoDB, Ruby on Rails, Redis, AWS

Projects

Chinchilla-Optimal Transformer Pre-training for Music — Fall 2025
Independent ML Systems Project | GitHub

- Independently designed and trained decoder-only Transformer models (NanoGPT) on the Lakh MIDI Dataset, scaling from 1M to 86M parameters using Chinchilla-optimal token budgeting ($D \approx 200N$) to study compute-efficient pretraining
- Optimized training throughput on NVIDIA H200 GPUs using BFloat16 mixed precision, Flash Attention, and `torch.compile`, achieving significantly improved training speed and sample efficiency over LSTM baselines
- Achieved a test perplexity of 2.20 with 100% syntactically valid output, generating coherent conditional and unconditional polyphonic music via Top-K sampling (K=600)
- **Tech Stack:** Python, PyTorch, CUDA, Flash Attention