Shriyansh Singh

+1-930-333-5141 | **shriyansh.singh24@gmail.com** | **LinkedIn**

SUMMARY

ML Engineer with expertise in **distributed systems** and **parallel computing**, seeking to pioneer innovative machine learning systems that scale efficiently and inspire creativity

EDUCATION

Indiana University Bloomington — Aug 2023 – May 2025
Master of Science in Data Science — *Indiana, United States*

PROFESSIONAL EXPERIENCE

Machine Learning Systems Engineer Intern — *April 2024 - Dec 2024*
Hyphenova AI — *Los Angeles, CA*

- **Architected** a **distributed computing** pipeline for LLM inference that reduced latency by 40% through efficient task scheduling and parallel request processing using **C++** and CUDA
- **Engineered** a fault-tolerant model serving system that handled 5,000+ concurrent requests with automatic failover mechanisms, ensuring 99.9% uptime for production services
- **Optimized GPU memory** utilization for LLaMA 2 model inference by implementing dynamic batch processing, increasing throughput by 30% without additional hardware
- **Streamlined** model deployment workflow with containerization using Docker and Kubernetes, reducing deployment time from days to hours while maintaining consistent testing environments
- **Integrated** comprehensive monitoring and logging infrastructure using Prometheus and Grafana that enabled real-time performance analysis and proactive system maintenance

Machine Learning Infrastructure Intern — *May 2022 - Oct 2022*
Enterprise Business Technologies Pvt Ltd — *Mumbai, India*

- **Developed** a **parallel computing** framework for text processing that scaled effectively across multiple nodes, reducing processing time for large document batches by 65%
- **Implemented** efficient **memory management** strategies for transformer model inference, allowing deployment on edge devices with limited resources
- **Designed** a modular task scheduling system in **Python** and **C++** that dynamically allocated computing resources based on workload priority and available hardware
- **Enhanced** model serving infrastructure by introducing asynchronous processing patterns, increasing system throughput while maintaining response time SLAs

PROJECTS

Distributed ML Training System | *C++, CUDA, PyTorch, MPI* — *Jan 2024 – Apr 2024*

- **Engineered** a distributed training framework for large language models that efficiently scales across multiple **GPU** nodes using parameter server architecture
- **Implemented** advanced **gradient compression** techniques that reduced network bandwidth requirements by 75% while maintaining model convergence characteristics
- **Designed** fault-tolerant checkpointing mechanisms that allow training to resume from hardware failures with minimal data loss, crucial for long-running training jobs
- **Integrated** adaptive learning rate scheduling based on compute node availability, optimizing training throughput in dynamic cluster environments

High-Performance Model Serving Platform | *Go, C++, RDMA, TensorRT* — *Sep 2023 – Dec 2023*

- **Created** a high-throughput model serving system utilizing **RDMA** networking for ultra-low latency inference in mission-critical applications
- **Optimized** transformer model inference through TensorRT integration and custom CUDA kernels, achieving 3.5x speedup compared to standard PyTorch deployment
- **Developed** an intelligent request batching system that maximizes GPU utilization while maintaining strict latency requirements for different service tiers
- **Implemented** comprehensive monitoring and automatic scaling capabilities that respond to traffic patterns, ensuring consistent performance under variable load

SKILLS & CERTIFICATIONS

Programming: C/C++, **Python**, **Go**, CUDA, Shell Scripting
Systems: Distributed Computing, **Parallel Processing**, **Task Scheduling**, **High-Performance Computing**
ML Infrastructure: PyTorch, TensorFlow, **GPU Optimization**, TensorRT, ONNX Runtime, **Model Serving**
Technologies: RDMA, Docker, Kubernetes, Prometheus, Linux, **Memory Management**, Cloud Computing