Satya Nandivada
207-252-2806 | sreesatyanandivada@gmail.com | Linkedin | Github | Website
Deep Learning Engineer | Model Optimization | GPU Systems | PyTorch | CUDA | Triton | TensorRT
Engineer specializing in optimizing large-scale generative and graph-based models through GPU-efficient architectures, dynamic compilation, and automated inference pipelines. Experienced in PyTorch, CUDA/Triton kernel profiling, and scalable ML infrastructure across AWS/GCP.

EDUCATION

Northeastern University
Master's in Electrical and Computer engineering *January 2023 - May 2025*
Gitam University,
Bachelors in Electronics and Communication Engineering *March 2017 – April 2021*

TECHNICAL SKILLS

Deep Learning Frameworks: PyTorch, HuggingFace, TensorRT, TorchDynamo, Triton, CUDA, CUTLASS
Optimization Techniques: Pruning, Quantization, Sparsity, Mixed Precision, Tensor Parallelism, Model Sharding
Programming & Systems: Python, C++, Docker, Kubernetes, AWS, GCP, Nextflow
Profiling & Performance Tools: Nsight Systems, PyTorch Profiler, CUDA Graphs, Torch Compile Stack
Other: Graph Neural Networks, Diffusion Models, ODE Solvers, Large Language Models

EXPERIENCE

DeepOMAP L.L.C, Portland, Maine
Founding Engineer Deep learning | Computational Biologist (Volunteer) *March 2025 - Present*

- Designed and optimized **multi-omics generative models** (GAT + Transformer fusion) using **PyTorch**, achieving **2× faster inference** via **GPU caching and mixed precision quantization**.
- Leveraged **TorchDynamo** and torch.compile() for dynamic graph tracing, improving runtime efficiency and enabling export for TensorRT-based deployment.
- Implemented automated **inference profiling and pruning routines** with CUDA-level hooks to monitor kernel execution and optimize throughput.
- Built modular inference pipelines integrated with **Nextflow**, **Docker**, and **AWS/GCP CI/CD**, enabling reproducible deployment of model optimization workflows.

Vanaja Systems Biology lab, Northeastern University
Deep learning Computational Biology Research Assistant/ Fall-Coop (2024) *March 2023 – Present*

- Trained and evaluated **ODE-based and neural hybrid models** for pathway simulation and drug-response prediction with optimized GPU utilization
- Integrated **custom CUDA kernels and PyTorch JIT** for differential equation solvers, improving parameter fitting efficiency by 45%
- Applied **model sparsification and quantization** to reduce computation time while maintaining >90% accuracy in biological simulations.

Gamer Alliance Members of Esports (GAME LLC) – Remote
Full-Stack AI Engineer *Jul 2025 – Present*

- Engineered scalable **AWS EC2 inference microservices** with **Flask + Triton Inference Server**, supporting >2000 concurrent sessions at <200 ms latency.
- Implemented **asynchronous model serving** and **TensorRT acceleration**, improving model throughput by 4×
- Automated profiling and performance diagnostics for distributed inference systems, leveraging GPU metrics and memory optimization strategies.
- Automated performance diagnostics across distributed microservices, reducing critical downtime incidents by **70%**.

Conduent Inc, Hyderabad, India
Application Developer-1, Full-Time *July 2021 – December 2022*

- Developed **production-grade distributed systems** using **Spring Boot, Spark, and Kubernetes**, emphasizing **low-latency, fault-tolerant AI pipelines** and optimized resource scaling (3.5× throughput boost).

CONFERENCES

- **Poster and Flash Talk Presentation:** A complete mathematical model of the MAPK pathway predicts mechanisms of resistance to BRAF inhibition in BRAF V600E-driven tumors. International Conference of Systems Biology (**ICSB 2024**), **IIT Mumbai**, India. December 2024.