KRISHNA PRANEET GUDIPATY

(+1) 425-542-4828 | kgudipaty@umass.edu | https://krishna-praneet.github.io/

RESEARCH INTERESTS

Distributed Systems, Operating Systems, Machine Learning, AI/ML Model Serving
Other interests: Quantitative methods for Finance, Options, Quantum Computing

EDUCATION

University of Massachusetts Amherst (UMass) — May 2027 (expected)
Ph.D. in Computer Science — GPA: 4.0/4.0
Advised by: Prashant Shenoy

University of Massachusetts Amherst (UMass) — May 2024
M.S. in Computer Science — GPA: 3.97/4.0
Master's Thesis advised by: Stefan Krastanov

Indian Institute of Technology Madras (IIT-M) — Dec 2020
B.Tech. in Metallurgical and Materials Engineering — GPA: 3.5/4.0
Bachelor's Thesis advised by: Satyesh Kumar Yadav

PUBLICATIONS

[1] **Gudipaty, Krishna Praneet**, W. A. Hanafy, L. Wu, *et al.*, "Practical considerations for failure resilient ml systems at the edge," in *MILCOM 2025 - 2025 IEEE Military Communications Conference (MILCOM) (to appear)*, 2025.

[2] **Gudipaty, Krishna Praneet**, W. A. Hanafy, K. Ozkara, *et al.*, "Mel: Multi-level ensemble learning for resource-constrained environments," *arXiv preprint arXiv:2506.20094*, 2025.

[3] A. Micciche, **Gudipaty, Krishna Praneet**, and S. Krastanov, "Quantum ldpc error correcting codes for use on 1d quantum dot arrays," in *APS March Meeting Abstracts*, vol. 2024, 2024, S46–010.

RESEARCH EXPERIENCE

Structural Resilience using Foundation Models (FMs) — Amherst, Mass.
Supervisors: Prashant Shenoy (UMass), Tarek Abdelzaher (UIUC) — Sep 2025 - Present

· Working on resource-constrained AI inference utilizing backbone-decoder architecture in FMs and through a proposed novel online fine-tuning and in-context learning approach to optimize failover time.

Practical considerations for AI model deployment — Amherst, Mass.
Supervisors: Prashant Shenoy (UMass) — Jun 2025 - Aug 2025

· Demonstrated practical considerations for AI model serving across enterprise edge (PowerEdge R360+NVIDIA A2s) and adversarial IoT (NVIDIA Jetson Orin NX) conditions for fault-tolerant model inference systems. Quantitatively measured the impact on application latency and accuracy by Model Loading times, Failure detection & Reconfiguration times, Mean Time To Recovery (MTTR), and the effect of co-location.

Multi-level Ensemble Learning — Amherst, Mass.
Supervisors: Prashant Shenoy (UMass), Suhas Diggavi (UCLA) — Nov 2024 - May 2025

· Introduced a novel ML training and architectural framework using a modified multi-objective loss for enhancing fault tolerance in resource-constrained edge AI environments that preserved up to **95.6%** accuracy using ensembles sized at **40%** of the original model. Attained **25%** lower inference latency through parallelized execution across edge servers compared to alternative baselines and across vision, audio, and NLP tasks.

Quantum LDPC Decoders Amherst, Mass.
Supervisors: Stefan Krastanov (UMass) Sep 2023 - May 2024

· Built the first Julia-native low-density parity-check (LDPC) decoders library with **500+** downloads on JuliaHub, including the first available open-source SOTA iterative decoder. Achieved **30%** improvement in decoding latency compared to existing baselines like *python-ldpc*, through optimized memory allocations and variable type profiling, and a user-friendly API for benchmarking qLDPC codes with just <**10** lines of code.

Machine Learning for Platinum Nanoclusters Chennai, India
Supervisors: Satyesh Kumar Yadav (IIT-M) Aug 2019 - May 2020

· Achieved up to **99.5%** accuracy in predictions for over 6 configurations for Platinum nanoclusters in various environments, and a computational speedup of **40%** on an IBM cluster through non-linear Machine Learning based frameworks for accelerating Ab-initio Molecular Dynamics simulations. Developed a numerical fingerprinting algorithm to capture translation/rotation invariance in 3-D and to construct feature vectors.

Assisted Defect Recognition with MATLAB Bengaluru, India
Supervisors: Jyani Vaddi (Boeing), Om Prakash (Boeing) Jun 2018 - Jul 2018

· Developed Image Processing Algorithms to identify the volume of shrinkage defects with a mean accuracy of **95%** in the torque arm of aircraft landing gear for Non-Destructive Evaluation. Analyzed raw DICOM images using a combination of mean and Canny Gaussian filter using DFT and documented an approach for defect segmentation by BPHF and difference of gaussians, followed by a routine of image dilation and erosion.

PROFESSIONAL EXPERIENCE

Center for Quantum Networks (CQN) Amherst, Mass.
API Developer Jan 2024 - May 2024

· Designed a REST API server utilizing Oxygen.jl framework to enable programmatic access to a quantum network simulator. Integrated HTTP endpoints for manipulating qubit registers and gates across nodes in the network using QuantumSavory.jl to enable reproducible research in quantum coding theory.

Deskera Singapore
Software Development Engineer Dec 2020 - July 2022

· Launched a Java microservices platform to reduce manual business management efforts by 40% through streamlining logistics across e-commerce portals and Point-of-Sales (POS) devices. Developed real-time CDC pipelines and webhooks to automate data reconciliation using Debezium and Kafka queues, reducing data propagation latency from ∼**60** minutes to <**5** seconds. Designed UI dashboards in React to provide users with no-code business intelligence, key metrics tracking, and integrating inventory synchronization.

Publicis Sapient Bengaluru, India
Software Development Engineer May 2019 - July 2019

· Refactored and delivered trading platform for short-term maturity instruments with RESTful microservices architecture using Java, improving order book latency by **25%** and reducing server utilization by **30%**. Revamped UI to a Single Page Architecture (SPA) with React and Bootstrap for real-time asset transactions and portfolio management, resulting in better page-to-page navigation and loading times by ∼**20%** on average.

SKILLS

Programming Languages	Python, Java, C/C++, Julia, MATLAB
Machine Learning Tools	Pytorch, Tensorflow, Scikit-learn, Pandas, Numpy, Spark
Database	SQL, MySQL, MongoDB, PostgreSQL, Redis, S3
Deployment & Monitoring	AWS, Google Cloud Platform, Jenkins, Docker, Kubernetes, Grafana
Web Development	HTML, CSS, JavaScript, React.js, Bootstrap, Swagger, Postman