Hema Venkata **Krishna** Giri Narra

+1-323-637-8771 | narrakrishnagiri@gmail.com | https://krishnagirinarra.github.io/

Career Summary

Currently working as a software engineer at Google. Received his Ph.D. working on improving the performance and privacy of machine learning systems. Has published in the top machine learning and high-performance computing conferences (NeurIPS, ICML, SC). Finalist for the best paper at SC. Has prior working experience in Samsung and Intel.

Education

University of Southern California, Los Angeles, USA **(2013-2020)**

Ph.D., Computer Engineering

Advisor: Prof. Murali Annavaram

M.S., Computer Science

Indian Institute of Technology (IIT) Madras, Chennai, India **(2006-2011)**

Bachelor of Technology and Master of Technology, Electrical Engineering

Languages, Frameworks, Tools, Environments

Python, PyTorch, Keras, TensorFlow, C, C++, Intel SGX, Docker, Kubernetes, Git, MPI, HTML, SQL, Verilog, Unix.

Research Projects

Origami Inference: Private Inference Using Hardware Enclaves

- Designed and implemented the Origami framework that partitions DNN models between secure hardware enclaves and unsecure accelerators. Designed GAN models for measuring privacy of the partitioned models.
- Implemented and trained conditional GAN models using Keras library. Used the SGXDNN library in C++ to execute CNN models on SGX enclaves.
- Evaluated the Origami framework on Intel SGX enclaves and demonstrated that up to 15x speedup in inference latency could be achieved compared to running the full model inside the secure SGX enclave.

Collage Inference: Using Coded Redundancy for Lowering Variance of Distributed Image Classification Systems

- Proposed collage-cnn models that perform multi-image classification and used them as low-cost redundancies to mitigate slowdowns in distributed inference systems.
- Implemented and trained the collage-cnn models using PyTorch and Darknet frameworks. Used the Flask web application framework in Python to implement the distributed inference system in the cloud.
- Demonstrated that deploying collage-cnn models can reduce the variance in latency by up to 15x, and the 99th percentile latency by up to 2x compared to alternate approaches.

S2C2: Slack Squeeze Coded Computing for Adaptive Straggler Mitigation

- Proposed S2C2, which exploits the data redundancy available in coded data and elastically distributes work based on speeds predicted from an LSTM model.
- Implemented the S2C2 code to train SVM and Logistic Regression models using numpy and xmlrpclib libraries in Python. Containerized the code using Docker and used Kubernetes to bootstrap a cluster of 50 nodes in the cloud and deploy the containers to them.
- S2C2 reduced the total compute latency by up to 39.3% over the conventional coded computation and by up to 19% over fine-grained replication.

GradiVeQ: Vector Quantization for Bandwidth-Efficient Gradient Aggregation in Distributed CNN Training

- Proposed GradiVeQ, a novel gradient compression technique that can significantly reduce the communication load in distributed CNN training.
- Implemented GradiVeQ using TensorFlow and mpi4py library in Python.
- GradiVeQ reduces the end-to-end training time by 4x over the uncompressed method, and by 1.4x over the 4-bit-QSGD method.

Industry Experience

Software Engineer, Google, Sunnyvale**,** CA **(Sep'20-Present)**

- Developing the infrastructure and methodology to analyze and optimize the performance of platforms.
- Developing the methodology to project the performance of future CPU server platforms.

Summer Intern, Datacenter Solutions Lab, **Samsung**, San Jose, CA **(May'16-Aug'16)**

- Developed a simulator to model the firmware and architecture of key-value SSDs.
- Demonstrated that using key-value SSDs can reduce write amplification by 15%.

Architecture Researcher Intern, SanDisk, San Jose, CA **(Jun'14-Aug'14)**

- Used storage runtime traces to analyze and compare the architectures of SanDisk SSDs and competitor SSDs.
- Proposed architectural changes to increase the performance of read operation on the SSDs.

Graphics Hardware Engineer, Intel, Bangalore, India **(Aug'11-Jul'13)**

- Worked on the design of logic blocks at the interface of GPU with the rest of the system-on-chip.
- Designed the logic blocks to handle data transfer across multiple voltages and frequency domains.

Selected Course work

• Machine Learning	• Distributed Systems	• Real-time Systems
• Database Systems	• Analysis of Algorithms	• Computer Systems Architecture
• Operating Systems	• Advanced Operating Systems	• Computer Networks

Selected Publications

1. **Krishna Narra** et al. "Origami Inference: Private Inference Using Hardware Enclaves". Published at IEEE Cloud 2021.
2. **Krishna Narra** et al. "Collage Inference: Using Coded Redundancy for Lowering Latency Variation in Distributed Image Classification Systems". Published at ICDCS 2020.
3. **Krishna Narra** et al. "Slack Squeeze Coded Computing for Adaptive Straggler Mitigation". Published at IEEE/ACM Supercomputing Conference (SC 2019). **Best paper finalist. Best student paper finalist.**
4. **Krishna Narra** et al. "Collage Inference: Achieving low tail latency during distributed image classification using coded redundancy models". CodML workshop at ICML 2019.
5. Mingchao Yu, Zhifeng Lin, **Krishna Narra** et al. "GradiVeQ: Vector Quantization for Bandwidth-Efficient Gradient Aggregation in Distributed CNN Training". Published at NeurIPS 2018.
6. Gunjae Koo, Kiran Matam, Te I, **Krishna Narra** et al. "Summarizer: Trading communication with computing near storage". Published at IEEE International Symposium on Microarchitecture (MICRO 2017).
7. L.Srivani, N.H.V.**Krishna Giri**, Shankar Ganesh, V.Kamakoti, "Generating Synthetic Benchmark Circuits for Accelerated Life Testing of Field Programmable Gate Arrays using Genetic Algorithm and Particle Swarm Optimization", Applied Soft Computing, Volume 27, February 2015.

Publications (Under submission)

1. Zhifeng Lin, **Krishna Narra** et al. "Train Where the Data is: A Case for Bandwidth Efficient Coded Training. ARXIV, abs/1910.10283, 2019.

Teaching Experience

Teaching Assistant for Computer Systems Architecture **(Spring 2017)**

Teaching Assistant for Advanced Topics in Microarchitecture **(Fall 2016)**

Teaching Assistant for Computer Networks **(Spring 2016)**

Teaching Assistant for Analysis of Algorithms **(Fall 2015, Summer 2015)**

Awards and Honors

- Best Dissertation award in Computer engineering at USC **(2019)**
- Best paper finalist, Best student paper finalist at the Supercomputing Conference (SC). **(2019)**
- USC Provost Ph.D. fellowship. **(2013-2017)**
- Instant recognition award for excellent service at Intel India. **(2012)**
- Recipient of M. J. Rao scholarship for excellent academic record in high school. **(2006-2011)**
- All India rank 606 out of 300,000 students in IIT JEE (Joint Entrance Examination). **(2006)**