Sandeep Polisetty

25 High street, Unit 3, Greenfield MA-01301.
☎ 413-801-3855 ✉ spolisetty@umass.edu Linkedin Google Scholar

Education

University of Massachusetts Amherst Aug 2020 – Present
Ph.D., Computer Science *Amherst, MA*

University of Massachusetts Amherst Aug 2016 – May 2020
M.S, Computer Science *Amherst, MA*

Indian Institute of Technology Sep 2008 – May 2013
B.Tech, M.Tech, Naval Architecture *Kharagpur, India*

Goals

I aspire to build infrastructure that is fueling the AI revolution. Efficient and easy to adopt frameworks are especially important with rising hardware costs. To support this goals, I have have worked at various levels across the framework such as efficient kernel design, parallelization strategies, compiler abstractions and hardware aware models to optimize performance.

Selected Publications

GSplit: Scaling graph neural network training on large graphs via split-parallelism under submission
S Polisetty, J Liu, K Falus, YR Fung, SH Lim, H Guan, M Serafini

Accelerating graph sampling for graph machine learning using GPUs Eurosys
A Jangda, S Polisetty, A Guha, M Serafini *2021*

Graphmini: Accelerating graph pattern matching using auxiliary graphs PACT
J Liu, S Polisetty, H Guan, M Serafini *2023*

Work Experience

Intel May 2024 – August 2024
AI Frameworks Engineer Intern *Santa Clara, CA*

- Worked on the torch extensions for large language models using inductor framework to utilize hardware optimized frameworks such as OneDNN and XeTLA
- Benchmarked the performance of quantization optimizations on important families of GenAI models such as Llama and Qwen

Reservoir Labs (acquired by Qualcomm) May 2021 – August 2021
Research Engineer Intern *Remote*

- Extended TVM intermediate representation Relay to support dynamic operators (eg. embedding bag) and generate c-code for downstream polyhedral compiler.
- Wrote transformation passes such as shape propagation and affine loop identification to increase the scope for polyhedral optimizations which are restricted to affine loops

Amazon May 2018 – August 2018
Software Engineer Intern *Seattle,WA*

- Extended in-house consistency service, to support zookeeper clients while maintaining original consistency protocol.
- Documented proof of correctness and passed the entire test suite leading to improved adoption by end users.

Research Experience

Distributed graph machine learning
advised by Prof. Marco Serafini and Prof. Hui Guan January 2021 - Present

- Designed a novel distributed sampling and training paradigm to eliminate redundancy across multiple mini-batches in GNN training across multiple GPUs
- Designed efficient abstractions to minimally change existing training and sampling loops. Light weight novel data structures and optimal CUDA kernels where designed to perform inter-GPU communication and co-ordination efficiently.
- Performed extensive experiments across different model architectures and hardware configurations. Our approach allows for deeper GNN models and outperforms the state of the art benchmarks such as Quiver and DGL upto **2x**. (work under submission)

Large Scale Semantic Scholar Paper Recommendation
at John Hopkins Speech and Language Workshop June 2023 - Aug 2023

- Processed massive semantic scholar database consisting of all published research, utilizing highly parallel cluster computing to create datasets consumable by deep learning frameworks.
- Used temporal partitioning of the graph and unified virtual adressing to train an extremely large graph containing over 500 GB of data on commodity V100 GPUs.
- Decreased training time further by using edge selection to limit training edges. The resulting GAT model delivered competitive accuracy relative to state of the art techniques based on matrix factorization, as well as decreasing training time to under a day from one week. (currently under submission)

GPU sampling for GNN
advised by Prof. Marco Serafini September 2020 - March 2021

- Integrated a novel in-GPU transit node-centric sampler called Nextdoor across various state-of-the-art graph training pipelines using torch extension to demonstrate the robustness of the designed abstraction and end to end training performance improvements (Published at Eurosys - 2021)

Efficient Graph Mining
advised by Prof. Marco Serafini November 2020

- Worked on eliminating redundancy in graph pattern mining by constructing light weight intermediate structures. Modified existing graph pattern matching algorithms where modified to utilize these light weight structures resulting in speedups upto **2x** over the state of the art.(Published at PACT-2023)

Technical Skills

Languages: C/C++, CUDA, python, java
Frameworks: tensorflow, pytorch, DGL, TVM, pybind, NSightSystems

Services

- Reviewer SC-2024
- Artifact Evaluater OSDI-2024, ATC-2024
- Artifact Evaluater OSDI-2020