Israt J Nisa

Santa Clara, CA • isratnisa295@gmail.com • https://isratnisa.github.io/• Phone: (419)764-0984

RESEARCH INTEREST

High-performance sparse matrix, tensor, and graph computations on parallel architectures, commonly applied in deep learning models, linear algebra, and computational genomics.

EDUCATION

Ph.D. in Computer Science and Engineering — Aug 2014 - Aug 2019
The Ohio State University - Columbus, OH
Supervisor: Prof. P. (Saday) Sadayappan
Thesis title: Architecture-aware Algorithm Design of Sparse Tensor/Matrix Primitives for GPUs

B.Sc. in Computer Science and Engineering — April 2007 - May 2011
University of Dhaka - Dhaka, Bangladesh

PROFESSIONAL EXPERIENCE

Amazon Web Service, Santa Clara, CA

Applied Scientist II — Nov 2023 – Present

- Optimize large language models for AWS's AI accelerator Trainium/Inferentia, which involves low-level system optimization and analysis of deep learning compilers and code generators.

Dec 2022 – Nov 2023

- Core development of AWS's open-source Graph Neural Network (GNN) framework - DGL (Deep Graph Library). Focused on system-level optimization of sparse and dense operators for GPUs, CPU-GPU communication, latency, and interactions with PyTorch backend during training.
- Designed and led the project to add the dgl.sparse library, which allows users to write GNN models using linear algebraic notation in addition to traditional message-passing APIs. Project contribution was interrupted due to team restructuring.

Applied Scientist I — Feb 2021 – Dec 2022

- Scaled billion-scale graphs via multi-GPU and multi-node training infrastructure. Reduced communication volume and developed efficient data-loading schemas.
- Conducted research on compressing large embedding tables for GPUs (KDD'22), temporal GNN (VLDB'22), scale GNN framework to handle billion scale graphs (KDD'23)

Berkeley Lab — Lawrence Berkeley National Laboratory, Berkeley, CA

Postdoctoral Fellow - Computational Research Division Jan 2020 – Jan 2021

- Worked on exascale solutions for microbiome analysis for GPUs.
- Implemented dynamic runtime fusion of parallel operators for graph analytics and machine learning.

NVIDIA Corporation, Santa Clara, CA

Deep Learning Software Intern May 2018 - June 2018

- Optimized the inference computation of deep learning algorithms.
- Implemented Google's Neural Machine Translation (GNMT) System using NVIDIA's TensorRT APIs
- Internship was interrupted due to visa issues

Pacific Northwest National Laboratory (PNNL), Richland, WA

PhD intern May 2017 - Nov 2017

- Application of neural network models to solve several classic sparse matrix related problems on GPUs
- Network models like Convolutional Neural Network (CNN) and Multiplayer Perceptron (MLP) are applied using via Google TensorFlow with GPU backend

Samsung Smart City Campus, Gumi, South Korea

Software Engineer Mar 2013 - May 2013

- Implementation of 3G/LTE protocols (3G-324M, H.263, H.245) at the framework and application layer for Samsung smart phones (based on C/C++, Android platform)
- Research and analysis on VoLTE stack, GSM/EDGE and UMTS/HSPA network

Samsung Institute of R&D Bangladesh, Dhaka, Bangladesh

Software Engineer Dec 2011 - July 2013

- Protocol analysis and development of the video call application for the Samsung smart phones
- Error handling in client and server side during video call establishment on Samsung Galaxy Note II

SELECTED PUBLICATIONS

Google scholar link: (https://scholar.google.com/citations?user=dfsJMOYAAAAJ&hl=en&oi=ao)

1. J Zhang, D Zheng, X Song, T Vasiloudis**, I Nisa**, J Lu: GraphStorm an Easy-to-use and Scalable Graph Neural Network Framework: From Beginners to Heroes, ACM SIGKDD'2023

2. **I Nisa**, M Wang, D Zheng, Q Fu, Ü Çatalyürek, G Karypis: Optimizing Irregular Dense Operators of Heterogeneous GNN Models on GPU, IPDPSW'2023

3. C Yin, D Zheng, **I Nisa**, C Faloutsos, G Karypis, R Vuduc: Nimble GNN Embedding with Tensor-Train Decomposition. ACM SIGKDD'2022

4. H Zhou, D Zheng, **I Nisa**, V Ioannidis, X Song, G Karypis: TGL: A General Framework for Temporal GNN Training on Billion-Scale Graphs. PVLDB'2022

5. **I Nisa,** Prashant Pandey, Marquita Ellis, Leonid Oliker, Aydin Buluç, Katherine Yelick: Distributed-Memory k-mer Counting on GPUs. IEEE International Parallel & Distributed Processing Symposium (IPDPS)'2021

6. **I Nisa**, J Li, A Sukumaran-Rajam, P Rawat, S Krishnamoorthy, P. Sadayappan: An Efficient Mixed-Mode Representation of Sparse Tensors. ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC)'2019

7. **I Nisa**, J Li, A Sukumaran-Rajam, R Vuduc, P Sadayappan: Load-balanced sparse MTTKRP on GPUs. IEEE International Parallel & Distributed Processing Symposium (IPDPS), 2019

8. C Hong, A. Sukumaran-Rajam, **I Nisa**, K Singh, P. Sadayappan: Adaptive Sparse Tiling for Sparse Matrix Multiplication. Principles and Practice of Parallel Programming (PPoPP), 2019

9. **I Nisa**, A Sukumaran-Rajam, SE Kurt, C Hong, P Sadayappan: Sampled Dense Matrix Multiplication for High-Performance Machine Learning. IEEE International Conference on High Performance Computing (HiPC), 2018 (best paper finalist)

10. G Moon, **I Nisa**, A Sukumaran-Rajam, B Bandyopadhyay, S Parthasarathy, P. Sadayappan: Parallel Latent Dirichlet Allocation on GPUs. International Conference on Computational Science (ICCS), 2018

11. **I Nisa**, A Sukumaran-Rajam, R Kunchum, P. Sadayappan: Parallel CCD++ on GPU for Matrix Factorization. Proceedings of the General Purpose GPUs (GPGPU), 2017

AWARDS

- Rising Stars in Computational and Data Sciences, 2020 (https://risingstars.oden.utexas.edu)
- Best paper finalist in IEEE International Conference on High Performance Computing (HiPC), 2018

PROFESSIONAL SERVICE

- Co-chair for the SAGE-S (Science Accelerating Girls' Engagement in STEM) planning committee
- Served as PC member at (selected):
  - International Conference on High-Performance Computing, Networking, Storage, and Analysis (SC)'23
  - IEEE International Parallel & Distributed Processing Symposium (IPDPS)'24
  - International Conference on Parallel Processing (ICPP)' 2022
  - Neurips Workshop (GLFrontier)'23

LANGUAGE SKILLS

- CUDA, C++, OpenMP, Python, MPI