855 Piedmont Ave NE, APT 1F
Atlanta, GA 30308, USA
+1 (832) 571 3421
ankush@gatech.edu
ankushmandal19@gmail.com
Linkedin Profile
Homepage

Ankush Mandal

Research Interests

Parallel Computing, Parallel Randomized Algorithms for Big-Data, Compiler Optimizations, Performance Optimization of Approximate Algorithms on Modern Architectures (e.g. Multi-core, Many-core, SIMD, GPU processors), High Performance Libraries for Machine Learning Kernels

Education

2017–present **Doctor of Philosophy**, *Georgia Institute of Technology*, Atlanta, GA, USA, (expected July 2020).
Advisor: Vivek Sarkar, Habanero Extreme Scale Software Research Laboratory, Georgia Institute of Technology
Co-advisor: Anshumali Shrivastava, RUSHLab, Rice University
Major: Computer Science
PhD thesis
Title: Enabling Parallelism and Optimizations in Data Mining Algorithms in the Presence of Power-law Data
Committee: Vivek Sarkar (Chair), Hyesoon Kim, Santosh Pande, Anshumali Shrivastava, Richard Vuduc

2014–2017 **Master of Science**, *Rice University*, Houston, TX, USA.
Major: Computer Science.
Master's thesis
Title: Optimizing Convolutions in State-of-the-art Convolutional Neural Networks on Intel Xeon Phi
Committee: Vivek Sarkar (Chair), Rajkishore Barik, Keith D. Cooper, Anshumali Shrivastava

2008–2012 **Bachelor of Engineering**, *Jadavpur University*, Kolkata, India.
Major: Electronics and Telecommunication Engineering.

Work Experience

Aug, 2017 - Present **Research Assistant**, *Georgia Institute of Technology*, Atlanta, GA, USA.
Research projects:

- **Optimizing Word2Vec (word embedding method) on latest x86 CPUs with wide SIMD units**
  - Developed insights into performance issues in Stochastic Gradient Descent (SGD) inside Word2Vec
  - Our solution involved both compiler optimizations (static multi-version code generation with novel vector register blocking scheme) and algorithmic modifications (reduced computation and improved data locality).
  - Achieved 9.5× speedup on SGD and 2.5× speedup on training time over state-of-the-art methods with AVX-512 ISA on Intel® Xeon® Platinum 8280 CPU (Cascade Lake architecture)
- **Improving concurrency in approximate frequency estimation methods on modern GPUs**
  - Proposed new nested sketching strategy suitable for GPU-scale parallelism
  - Our approach exploits power-law behavior in data to reduce contention in atomic updates, gave detailed theoretical analysis
  - Attained throughput improvement of 32× over competing GPU-based method on nVidia® Tesla® V100 GPU and 272× over state-of-the-art sequential CPU-based method on Intel® Xeon® Platinum 8180 CPU (Skylake architecture)
- **Approximate K most frequent elements finding on massively parallel distributed + shared memory systems**
  - Tackled important data mining problem of finding TopK frequent items in distributed data streams
  - Combined sketch-based and counter-based approaches in unique way to aid parallelization while retaining fast update time
  - Implemented using MPI for multi-node parallelism and OpenMP for multi-core parallelism
  - Demonstrated 2.5× speedup over competing methods on clusters of Intel® Westmere and IBM Power®7 CPUs

Technology C, C++, OpenMP, MPI, CUDA

May, 2018 - July, 2018 **Intern for R&D of Energy and Performance Analysis**, *Intel*, Austin, TX, USA.
Mentor: David Kuck
- Performed energy and performance analysis of convolutions in popular Convolutional Neural Networks on x86 CPUs (Broadwell and Skylake architectures)
- Came up with performance-energy trade-off variation for direct convolution kernel when applying different compiler optimizations

Technology C, OpenMP

Aug, 2014 - Aug, 2017 **Research Assistant**, *Rice University*, Houston, TX, USA.
Research projects:
- Focused on improving performance of parallel machine learning algorithms
- Studied locality-sensitive hashing and heavy hitter detection on different architectures
- Worked on parallelizing forward-backward algorithm over profile-Hidden Markov Model in bioinformatics application.

Technology C, C++, OpenMP, CUDA

Jan, 2017 - May, 2017 **Graduate Intern**, *Intel Labs*, Santa Clara, CA, USA.
Mentor: Rajkishore Barik
- Optimized direct convolution kernel for convolutions in popular Convolutional Neural Networks on x86 CPUs targeting High-Performance Computing, specifically Intel Xeon Phi Knights Landing CPU.
- Contributed to open source LIBXSMM library
- Achieved ninja performance via JIT-based runtime code specialization and compiler optimizations
- Showed orders of magnitude performance improvement compared to popular matrix-multiplication (GEMM) based approach employing Intel® MKL

Technology C, OpenMP

June, 2016 - Aug, 2016 **Intern**, *AMD*, Austin, TX, USA.
Mentor: Mayank Daga
- Worked on in-house auto-tuning GEMM framework and Caffe (popular Deep Learning framework)
- Focused on analyzing and improving performance of auto-tuning GEMM framework for Caffe related problems on GPU architecture.
- Showed 5$\times$ performance improvement over ViennaCL for forward pass on convolution layers of AlexNet.

Technology C++, OpenCL

Aug, 2015 - May, 2016 **Teaching Assistant**, *Rice University*, Houston, TX, USA.
Courses:
- "Introduction to Computer Systems" COMP 321 - focuses on underlying aspects of computer systems
- "Parallel Computing" COMP 422 - introduction to foundations of parallel computing including the principles of parallel algorithm design, programming models for shared- and distributed-memory systems, parallel computer architectures

Technology C, C++, Cilk, OpenMP, Pthread, MPI, CUDA

Publications (selected)

Ankush Mandal, Anshumali Shrivastava, and Vivek Sarkar. Ninjavec: Learning word embeddings with word2vec at lightning speed. (In preparation — draft copy available on request).

Ankush Mandal, Anshumali Shrivastava, and Vivek Sarkar. Matryoshka: a nested sketching strategy for massive parallelism and skewed data. (In preparation — draft copy available on request).

Ankush Mandal, He Jiang, Anshumali Shrivastava, and Vivek Sarkar. Topkapi: parallel and fast sketches for finding top-k frequent elements. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10898–10908, 2018.

Ankush Mandal, Rajkishore Barik, and Vivek Sarkar. Using dynamic compilation to achieve ninja performance for cnn training on many-core processors. In *European Conference on Parallel Processing (Euro-Par)*, pages 265–278. Springer, 2018.

Swagatam Das, Ankush Mandal, and Rohan Mukherjee. An adaptive differential evolution algorithm for global optimization in dynamic environments. *IEEE Transactions on Cybernetics*, 44(6):966–978, 2013.

Courses Taken (selected)

ECE 8903 Special Problem, *by Tushar Krishna*.

Guided research project course. The objective of this project is to explore mappings of sparse and irregular matrix-multiplications on to spatial accelerators.

COMP 582: Graduate Design and Analysis of Algorithms, *by Krishna V. Palem*.

This course discusses methods for designing and analyzing computer algorithms and data structures. The focus of this course is on the theoretical and mathematical aspects of algorithms and data structures.

COMP 522: Multicore Computing, *by John Mellor-Crummey*.

Topics include multicore microprocessors, memory hierarchy (cache organization alternatives), multithreaded programming models, scheduling (including work stealing), memory models (which specify program behavior), synchronization (including wait-free synchronization), transactional memory (hardware and software), concurrent data structures, debugging (including race detection), and performance analysis.

Course Project: Exploring Heterogeneous C++ Compiler (HCC) with AMD A10 APU. In this project, I explored the use of the Heterogeneous C++ Compiler (HCC) on AMD A10 APU as target architecture. Experimented with parallelizing the LU Decomposition benchmark from the Rodinia suite using HCC.

COMP 640: Graduate Seminar in Machine Learning, *by Anshumali Shrivastava*.

This research seminar discusses recent advances and trends in machine learning. Topics include Support Vector Machines, Hashing, Random Projections, Latent Dirichlet Allocation, and Markov Chain Monte Carlo.

COMP 422: Intro to Parallel Computing, *by John Mellor-Crummey*.

This course introduces principles of parallel algorithm design, analytical modeling of parallel programs, and programming models for shared- and distributed-memory systems. Materials on emerging multicore hardware, shared-memory programming models, message passing programming models used for cluster computing, data-parallel programming models for GPUs, and problem-solving on large-scale clusters using MapReduce are also covered in this course.

COMP 412: Compiler Construction, *by Keith Cooper & Linda Torczon*.

This course provides an overview of the issues that arise in the design and construction of translators for programming languages. The course emphasizes techniques that have direct application to the development of compilers, including syntax analysis and simple program optimization.

COMP 440: Artificial Intelligence, *by Devika Subramanian*.

This is a foundational course in artificial intelligence, the discipline of designing intelligent agents. This course covers the design and analysis of agents that do the right thing in the face of limited information and computational resources. Tools from computer science, probability theory, and game theory are used.

Distinction and Awards

2018 **NeurIPS Travel Award**, NeurIPS Foundation.

Grant for attending NeurIPS'18 conference

2014 **CIDSE Doctoral Fellowship**, Arizona State University.

This prestigious fellowship is awarded to incoming graduate students for excellent academic accomplishments. (Declined award because I decided to join Rice University instead.)

2008 **Scholarship for excellent academic performance in Higher Secondary Examination (10+2 level)**, India.

Awarded scholarship under the Scheme of Scholarship from the Ministry of Human Resource Development (MHRD), Government of India

Personal Information

Visa Status: F1