Abhishek Roy

1040 N Pleasant St, 76, Amherst, MA 01002
+1 (413) 695 0872 • aroy@cs.umass.edu • cs.umass.edu/~aroy

Interests

- Database Systems
- Distributed and Parallel Databases
- Large-Scale Data Analysis
- Genomic Data Analysis

Education

University of Massachusetts Amherst, MS/PhD 2011–Present
- PhD candidate in Computer Science
- Advised by Prof. Yanlei Diao and Prof. Prashant Shenoy
- GPA: 3.9/4

Indian Institute of Technology (IIT) Guwahati, B.Tech 2005–2009
- Bachelor of Technology in Computer Science and Engineering
- Thesis: Verifying Correctness of Dynamic Software Transactional Memory
- Advised by Prof. Purandar Bhaduri

Publications

1. **Abhishek Roy**, Yanlei Diao, Toby Bloom, Uday Evani, and Clinton Howarth. Building and benchmarking a parallel deep analysis pipeline. *(submitted to a major database conference)*
2. Yanlei Diao, **Abhishek Roy**, and Toby Bloom. Building highly-optimized, low-latency pipelines for genomic data analysis. In *Proceedings of CIDR'15*, 2015. *(main student author)*
3. **Abhishek Roy**, Yanlei Diao, Evan Mauceli, Yiping Shen, and Bai-Lin Wu. Massive genomic data processing and deep analysis. *Proc. VLDB Endow.*, 5(12):1906–1909, August 2012.
4. Ling Chen and **Abhishek Roy**. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 523–532, New York, NY, USA, 2009. ACM.

Research Experience

University of Massachusetts Amherst, Research Assistant to Prof. Yanlei Diao 2011–Present
- Affiliated with Database and Information Management Lab
- Genomic Scalable Analysis with Low Latency (GESALL) project:
  - Improving the efficiency of Hadoop for complex analytics pipelines
  - Designed a distributed storage system for genomic data
    - Added logical partitioning and colocation features on top of HDFS
    - Developed a storage substrate to support binary, compressed and indexed data on HDFS
  - Built a parallel platform based on big data technology to speed up genomic data processing pipelines
    - Parallelized existing analysis programs using minimal MapReduce rounds
    - Implemented complex partitioning schemes
  - Working with New York Genome Center to guarantee the accuracy of results
  - Assembled and profiled the state-of-art pipeline for genomic data analysis
  - Expert in measuring performance of Hadoop-based systems and multi-threaded programs
- Previous work on association mining:
  - Used association mining to find the causal relationship between variants and phenotypes
  - Extended existing algorithms to support - a new interestingness metric, low support values, and proximity-aware mining

Industry Experience

Strand Life Sciences, Software Engineer 2009–2011
- Involved in the development of first version of Strand NGS, a desktop based software for next generation sequencing data analysis
- Implemented algorithms and statistical tests to analyze large datasets
- Contributed at every stage of product development life cycle

Teaching Experience

CMPSCI 445 (Undergraduate Database Systems), Teaching Assistant Spring 2014
- Designed and evaluated the course project
- Delivered introductory lectures on SQL and schema refinement

Internships

Boston Children's Hospital, Boston, Massachusetts Summer 2012
- Worked in Genetic Diagnostic Laboratory with Dr. Yiping Shen and Evan Mauceli

L3S Research Center, Hanover, Germany Summer 2008
- Supervised by Dr. Ling Chen and Prof. Wolfgang Nejdl
- Developed a wavelet based approach to detect events from Flickr tags

Embedded Systems Lab, IIT Delhi, India Summer 2007
- Supervised by Prof. Subrat Kar

Graduate Courses

Systems:		**Analysis:**	
Database Design and Implementation	A	Advanced Algorithms	A
Advanced Database Systems	A-	Artificial Intelligence	A-
Distributed Operating Systems	A	Probabilistic Graphical Models	A
Fast Algorithms and Parallel Processing	A	Mathematical Statistics I	A
Performance Evaluation	Audit	Integer Programming	Audit

Posters & Demonstrations

- Building Highly-Optimized, Low-Latency Pipelines for Genomic Data Analysis
  - New England Database Summit (NEDB), January 2015
- Massive Genomic Data Processing and Deep Analysis
  - UMass Big Data Workshop, October 2013
  - New England Database Summit (NEDB), January 2013
  - Very Large Data Bases (VLDB), September 2012

Skills

Hadoop: YARN, HDFS, MapReduce, HBase
Databases: PostgreSQL, MySQL, RocksDB
Programming: C, Java, Python, Go, MATLAB, AMPL, PHP, SQL, Assembly, Bash, Linux systems
Frameworks/Tools: CUDA, Weka, Socket programming, RPC, RMI, Spin Model Checker, Tangram2
Genomic toolkits: HTSJDK, PicardTools, GATK, Genome Browser
Performance debugging: gdb, perf, sar, SystemTap, JDK profilers, Valgrind

Other

- External Reviewer, ACM SIGMOD Record, 2015
- Recruitment Team, Strand Life Sciences, 2010