Uday Kumar Reddy Bondhugula

Professor
Department of Computer Science and Automation
Indian Institute of Science
Bengaluru, Karnataka 560012 INDIA

Tel: +91-80-2293-3249
Fax: +91-80-2360-2911
udayb@iisc.ac.in
http://www.csa.iisc.ac.in/∼udayb

Name as it appears on all publications: Uday Bondhugula

Research Interests

Compilation and parallelization for multicores, accelerators, and domain-specific hardware; high-performance domain-specific languages and compilers; automatic parallelization; polyhedral framework; MLIR. Domains of interest: high-performance AI, deep learning, stencils, and dense linear algebra.

Education

- Ph.D. in Computer Science and Engineering — Sep 2004 – Aug 2008
  The Ohio State University (OSU) — Columbus, OH, USA
  Thesis: *Effective Automatic Parallelization and Locality Optimization using the Polyhedral Framework*
  Advisor: Prof. P. Sadayappan

- Bachelor of Technology in Computer Science and Engineering — Jul 2000 – Jul 2004
  Indian Institute of Technology (IIT), Madras. — Chennai, India

Professional Experience

- Founder, CTO, and CEO — May 2019 – present
  PolyMage Labs — Bengaluru, India

- Professor, Mindtree Chair — Sep 2023 – present
  Department of Computer Science and Automation
  Indian Institute of Science — Bengaluru, India

- Associate Professor, Mindtree Chair — Dec 2016 – Aug 2020, May 2022 – Sep 2023
  Department of Computer Science and Automation
  Indian Institute of Science — Bengaluru, India
  (on leave of absence) Sep 2020 – Apr 2022

- Visiting Researcher — Mar 2018 – Mar 2019
  Google Brain team
  Google, Mountain View — California, USA

- Assistant Professor — Jan 2011 – Dec 2016
  Department of Computer Science and Automation
  Indian Institute of Science — Bengaluru, India

- Postdoctoral Research Scientist — Oct 2008 – Dec 2010
  Advanced Compiler Technologies
  IBM T.J. Watson Research Center, Yorktown Heights, New York — Yorktown Heights, NY, USA

- Visiting Researcher — Mar 2008 – May 2008
  ALCHEMY team
  INRIA Saclay (INRIA Futurs), Île-de-France — Orsay, France

- Research Intern   Jun 2007 – Sep 2007
  Advanced Compilation Technologies
  IBM T.J. Watson Research Center, Yorktown Heights, NY   Yorktown Heights, NY

- Graduate Research Associate   Apr'05 – Jun'07, Oct'07 – Aug'08
  Dept of Computer Science and Engineering
  The Ohio State University   Columbus, OH, USA

- Graduate Teaching Associate   Sep 2004 – Mar 2005
  Dept of Computer Science and Engineering, OSU.   Columbus, OH, USA

- Summer Intern   Jun 2003 – Aug 2003
  Trilogy Software Inc.   Bengaluru, India

Publications

Google Scholar profile, DBLP

1. SilvanForge: A Schedule-Guided Retargetable Compiler for Decision Tree Inference
   Ashwin Prasad, Sampath Rajendra, Kaushik Rajan, R Govindarajan, and Uday Bondhugula
   ACM SOSP 2024.

2. HIR: An MLIR-based Intermediate Representation for Hardware Accelerator Description
   Kingshuk Majumder and Uday Bondhugula
   International Conference on Architectural Support for Programming Languages and Operating Systems
   (ASPLOS), 13 pages, 2024 (to appear).

3. Treebeard: An Optimizing Compiler for Decision Tree-Based ML Inference
   Ashwin Prasad, Sampath Rajendra, Kaushik Rajan, R Govindarajan, Uday Bondhugula.
   IEEE/ACM International Symposium on Microarchitectures (MICRO), Oct 2022.

4. MLIR-Based Code Generation for GPU Tensor Cores
   Navdeep Katel, Vivek Khandelwal, and Uday Bondhugula.
   ACM/IEEE International conference on Compiler Construction (CC), Apr 2022.

5. A Practical Tile Size Selection Model for Affine Loop Nests.
   Kumudha Narasimhan, Aravind Acharya, Abhinav Baid, and Uday Bondhugula.
   ACM International Conference on Supercomputing (ICS'21), Jun 2021.

6. MLIR: Scaling Compiler Infrastructure for Domain-Specific Computation
   Chris Lattner, Mehdi Amini, Uday Bondhugula, Albert Cohen, Andy Davis, Jacques Pienaar, River
   Riddle, Tatiana Shpeisman, Nicolas Vasilache, and Oleksandr Zinenko.
   ACM International Symposium on Code Generation and Optimization (CGO), pages 2–14, Mar 2021.

7. An Effective Fusion and Tile Size Model for PolyMage
   Abhinav Jangda and Uday Bondhugula
   ACM Transactions on Programming Languages and Systems (TOPLAS), 42, 3, Article 12, 27 pages,
   Nov 2020.

8. Optimizing the Linear Fascicle Evaluation Algorithm for Multi-Core and Many-Core Systems
   Karan Aggarwal and Uday Bondhugula
   ACM Transactions on Parallel Computing, Article 22, Nov 2020.

9. Effective Loop Fusion in Polyhedral Compilation using Fusion Conflict Graphs
   Aravind Acharya, Uday Bondhugula, Albert Cohen.
   ACM Transactions on Architecture and Code Optimization (TACO), Article 26, Sep 2020.

10. Bitwidth Customization in Image Processing Pipelines using Interval Analysis and SMT Solvers
    Suresh Purini, Vinamra Benara, Ziaul Chowdhury, Uday Bondhugula
    ACM SIGPLAN International Conference on Compiler Construction (CC), Feb 2020.

11. Optimizing the Linear Fascicle Evaluation Algorithm for Many-Core Systems
    Karan Aggarwal, Uday Bondhugula
    International Conference on Supercomputing (ICS), Jun 2019.

12. Polyhedral Auto-Transformation with No Integer Linear Programming
    Aravind Acharya, Uday Bondhugula, Albert Cohen
    ACM SIGPLAN PLDI 2018.

13. An Effective Fusion and Tile Size Model for Optimizing Image Processing Pipelines
    Abhinav Jangda, Uday Bondhugula
    ACM SIGPLAN symposium on Principles and Practice of Parallel Programming (PPoPP), Feb 2018.
    Artifact evaluated (reusable and available).

14. Optimizing Geometric Multigrid Method Computation using a DSL Approach
    Vinay Vasista, Kumudha KN, Siddharth Bhat, Uday Bondhugula
    Supercomputing (SC), Nov 2017.

15. Diamond Tiling: Tiling Techniques to Maximize Parallelism for Stencil Computations
    Uday Bondhugula, Vinayaka Bandishti, Irshad Pananilath
    IEEE Transactions on Parallel and Distributed Systems (TPDS), pgs 1285–1298, vol 27, issue 3, May 2017.

16. A DSL Compiler for Accelerating Image Processing Pipelines on FPGAs
    Nitin Chugh, Vinay Vasista, Suresh Purini, Uday Bondhugula
    IEEE International conference on Parallel Architectures and Compilation Techniques (PACT 2016), Sep 2016.

17. Compiling Affine Loop Nests for a Dynamic Scheduling Runtime on Shared and Distributed Memory
    Roshan Dathathri, Ravi Teja Mullapudi, Uday Bondhugula
    ACM Transactions on Parallel Computing, volume 3, issue 2, Jul 2016.

18. SMO: An Integrated Approach to Intra-Array and Inter-Array Storage Optimization
    Somashekaracharya Bhaskaracharya, Uday Bondhugula, Albert Cohen
    ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL), Jan 2016.

19. The Pluto+ Algorithm: A Practical Approach for Parallelization and Locality Optimization of Affine Loop Nests
    Uday Bondhugula, Aravind Acharya, Albert Cohen
    ACM Transactions on Programming Languages and Systems, volume 38, issue 3, Apr 2016.

20. Automatic Storage Optimization for Arrays
    Somashekaracharya Bhaskaracharya, Uday Bondhugula, Albert Cohen
    ACM Transactions on Programming Languages and Systems (TOPLAS), volume 38, issue 3, Apr 2016.

21. An Optimizing Code Generator for a Class of Lattice-Boltzmann Computations
    Irshad Pananilath, Aravind Acharya, Vinay Vasista, Uday Bondhugula
    ACM Transactions on Architecture and Code Optimization (TACO), Volume 12 Issue 2, Article No. 14, Jul 2015.

22. PolyMage: Automatic Optimization for Image Processing Pipelines
    Ravi Teja Mullapudi, Vinay Vasista, Uday Bondhugula
    International Conference on Architecture Support for Programming Languages and Operating Systems (ASPLOS 2015), Mar 2015.

23. PLUTO+: Near-Complete Modeling of Affine Transformations for Parallelism and Locality
    Aravind Acharya, Uday Bondhugula
    ACM SIGPLAN PPoPP 2015, Feb 2015.

24. Tiling and Optimizing Time-Iterated Computations over Periodic Domains
    Uday Bondhugula, Vinayaka Bandishti, Albert Cohen, Guillain Potron, Nicolas Vasilache
    IEEE International conference on Parallel Architectures and Compilation Techniques (PACT 2014), Aug
    2014.
    Nominated for the best paper award.

25. Effective Automatic Computation Placement and Data allocation for Parallelization of Regular Programs
    Chandan Reddy, Uday Bondhugula
    ACM International Conference on Supercomputing (ICS), Jun 2014, Munich, Germany.

26. Automatic Data Allocation and Buffer Management for Multi-GPU Machines
    Thejas Ramashekar, Uday Bondhugula
    ACM Transactions on Architecture and Code Optimization (Nov 2013), selected for presentation at
    HiPEAC '14, Jan 2014, Vienna, Austria.

27. Compiling Affine Loop Nests for Distributed-Memory Parallel Architectures
    Uday Bondhugula
    ACM/IEEE Supercomputing (SC '13), Nov 2013.

28. Generating Efficient Data Movement Code for Heterogeneous Architectures with Distributed-Memory
    Roshan Dathathri, Chandan Reddy, Thejas Ramashekar, Uday Bondhugula
    International conference on Parallel Architectures and Compilation Techniques (PACT 2013), Sep 2013.

29. PolyGLoT: A Polyhedral Loop Transformation Framework for a Graphical Dataflow Language
    Somashekaracharya Bhaskaracharya, Uday Bondhugula
    International conference on Compiler Construction (CC 2013), Mar 2013, Rome, Italy.

30. Tiling Stencil Computations to Maximize Parallelism
    Vinayak Bandishti, Irshad Pananilath, and Uday Bondhugula
    ACM/IEEE Supercomputing, Nov 2012, Utah, USA.

31. Loop Transformations: Convexity, Pruning, and Optimization
    Louis-Noel Pouchet, Uday Bondhugula, Cedric Bastoul, Albert Cohen, J Ramanujam, P Sadayappan,
    and Nicolas Vasilache
    ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (PoPL), 2011.

32. Combined Iterative and Model-driven Optimization in an Automatic Parallelization Framework
    Louis-Noel Pouchet, Uday Bondhugula, Cedric Bastoul, Albert Cohen, J Ramanujam, P Sadayappan
    Supercomputing (SC) 2010.

33. A Model for Fusion and Code Motion in an Integrated Auto-Parallelizing Compiler
    Uday Bondhugula, Oktay Gunluk, Sanjeeb Dash, and L. Renganarayana
    International Conference on Parallel Architectures and Compilation Techniques (PACT), Sep 2010, Vi-
    enna, Austria.

34. Compact multi-dimensional kernel extraction for register tiling
    L. Renganarayana, Uday Bondhugula, Salem Derisavi, Alexandre E. Eichenberger, and Kevin O'Brien
    Supercomputing 2009

35. Compiler-Assisted Dynamic Scheduling for Effective Parallelization of Loop Nests on Multicore Proces-
    sors
    M. Baskaran, N. Vydyanathan, Uday Bondhugula, J. Ramanujam, A. Rountev, and P. Sadayappan.
    ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP'09), Feb 2009,
    Raleigh, North Carolina.

36. Data Layout Transformation for Enhancing Locality on NUCA Chip Multiprocessors
    Qingda Lu, Christophe Alias, Uday Bondhugula, Thomas Henretty, Sriram Krishnamoorthy, J. Ramanu-jam, Atanas Rountev, P. Sadayappan, Yongjian Chen, Haibo Lin, and Tin-fook Ngai.
    International Conference on Parallel Architectures and Compilation Techniques (PACT), 2009

37. A Practical Automatic Polyhedral Parallelizer and Locality Optimizer
    Uday Bondhugula, A. Hartono, J. Ramanujam, and P. Sadayappan.
    ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '08), Jun 2008, Tucson, Arizona.
    ACM SIGPLAN Most Influential Paper Award in 2018.

38. Automatic Transformations for Communication-Minimized Parallelization and Locality Optimization in the Polyhedral Model
    Uday Bondhugula, M. Baskaran, S. Krishnamoorthy, J. Ramanujam, A. Rountev, and P. Sadayappan.
    International Conference on Compiler Construction (ETAPS CC'08), Apr 2008, Budapest, Hungary.

39. A compiler framework for optimization of affine loop nests for GPGPUs
    M. Baskaran, Uday Bondhugula, S. Krishnamoorthy, J. Ramanujam, A. Rountev, and P. Sadayappan.
    ACM International Conference on Supercomputing (ICS'08), Jun 2008, Kos, Greece.

40. Automatic data movement and computation mapping for multi-level parallel architectures with explicitly managed memories
    M. Baskaran, Uday Bondhugula, S. Krishnamoorthy, J. Ramanujam, A. Rountev, and P. Sadayappan.
    ACM SIGPLAN PPoPP'08, Feb 2008, Salt Lake City, Utah.

41. Automatic mapping of nested loops to FPGAs
    Uday Bondhugula, J. Ramanujam, and P. Sadayappan.
    ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '07), Mar 2007, San Jose, California.

42. Effective automatic parallelization of stencil computations
    S. Krishnamoorthy, M. Baskaran, Uday Bondhugula, J. Ramanujam, A. Rountev, and P. Sadayappan.
    ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI '07), Jun 2007, San Diego, California.

43. Hardware/software integration for FPGA-based all-pairs shortest-paths
    Uday Bondhugula, A. Devulapalli, J. Dinan, J. Fernando, P. Wyckoff, E. Stahlberg, and P. Sadayappan.
    IEEE Symposium on Field-Programmable Custom Computing Machines (FCCM '06), Apr 2006, Napa Valley, California.

44. Parallel FPGA-based all-pairs shortest-paths in a directed graph
    Uday Bondhugula, Ananth Devulapalli, Joseph Fernando, Pete Wyckoff, and P. Sadayappan
    20th IEEE International Parallel and Distributed Processing Symposium (IPDPS '06), Apr 2006, Rhodes, Greece.

45. High performance RDMA-based all-to-all broadcast for InfiniBand clusters
    S. Sur, Uday Bondhugula, A. Mamidala, H.-W. Jin, and D. K. Panda
    12th IEEE International Conference on High Performance Computing (HiPC '05), Dec 2005.

Software and Tools

1. **MLIR**
   https://mlir.llvm.org/
   Founding team member of the MLIR project, and co-developer of the early infrastructure, especially, the polyhedral/mid-level analysis and optimization infrastructure; open-sourced by Google in Apr 2019 and now an LLVM sub-project with high traction in the industry/community.

The MLIR project was initiated to deliver the next generation optimizing compiler infrastructure with a focus on serving the computational demands of AI and machine learning programming models. At Google itself, one of the project's goals is to address the compiler challenges associated with the TensorFlow ecosystem. MLIR is a new intermediate representation designed to provide a unified, modular, and extensible infrastructure to progressively lower dataflow compute graphs, through loop nests potentially, to high-performance target-specific code. MLIR shares similarities with traditional CFG-based three-address SSA representations (including LLVM IR or Swift intermediate language), but also introduces notions from the polyhedral compiler framework as first class concepts to allow powerful analysis and transformation in the presence of loop nests and multi-dimensional arrays. MLIR supports multiple front- and back-ends and uses LLVM IR as one of its primary code generation targets. It is thus a very useful infrastructure for developing new compilers, especially to solve the compilation challenges involved in targeting emerging AI and machine learning programming languages/models to the plethora of specialized accelerator chips.

2. **Pluto**
   http://pluto-compiler.sourceforge.net
   Original and lead author of Pluto.

   Pluto is a source-to-source parallelization and optimization tool based on the polyhedral compiler framework. It can automatically optimize affine loop nests (sequences of imperfectly nested loops with regular data access patterns) for parallelism and locality using affine transformations. It can target both shared-memory multicore architectures (by generating code with OpenMP parallel pragmas) and distributed-memory architectures (by generating message passing MPI code). Pluto/Pluto+ is extensively used for advanced experimentation with loop optimization and parallelization, optimization of scientific stencil computations, and in university courses teaching loop transformations.

3. **PolyMage**
   http://mcl.csa.iisc.ernet.in/polymage.html

   PolyMage is a domain-specific language and compiler for automatic parallelization and optimization of image processing pipelines. PolyMage takes an image processing pipeline expressed by the user in a high-level language (embedded in Python) and generates a C++ implementation of the pipeline optimized using the polyhedral framework as the intermediate representation. It uses OpenCV for image I/O handling, islpy/ISL for integer set operations, 'cgen' for AST code generation and 'OpenMP' to mark parallel loops. PolyMage uses an asymmetric overlapped tiling technique (overlapped tiling extended for heterogeneous accesses and non-constant dependence vectors) to exploit locality and parallelism simultaneously. It uses a model-driven approach to automatically fuse image processing pipeline stages for tiling, and employs an in-built autotuner to find the best performing code within a small well-defined search space.

Awards and Honors

- Qualcomm Faculty Research Award 2022

- Awarded the Mindtree Chair position at the Department of CSA

- Honorable Mention - ACM India Early Career Research Award 2020

- Cray APJ Abdul Kalam HPC award 2019 — Young Researcher HPC Systems

- ACM SIGPLAN PLDI Most Influential Paper award in 2018 for PLDI 2008 paper

- ACM SIGPLAN PLDI 2017 Distinguished Reviewer Award as PC member

- Indian National Science Academy Medal for Young Scientists 2017

- Indian National Academy of Engineering Young Engineer Award 2016

- Awarded Indian Academy of Sciences Young Associate 2016–2019

- Google Faculty Research Award 2015

- Nominated for the best paper award at PACT 2014 for 'Tiling and Optimizing Time-Iterated Computations over Periodic Domains'

- INRIA Associate Team award (2013–2015) from INRIA, France (awarded on a competitive basis worldwide)

- Nominated for the ACM SIGPLAN doctoral dissertation award 2008

- ACM SIGPLAN student travel award for PLDI 2008

- All-India Rank 84 (top 0.06%) at the Indian Institutes of Technology Joint Entrance Examination 2000, out of a total of about 1,27,000 candidates

- Represented state of Andhra Pradesh, India at the Indian National Mathematical Olympiad in 1999

- Pratibha scholarship by the Govt of Andhra Pradesh (2000–2004) for performance at IIT-JEE 2000

- Recipient of the National Talent Search Exam (NTSE) scholarship (India) in 1998

Research Grants

- Department of Science and Technology / SERB extra-mural research grant 2017–2020

- Google Faculty Research award (2015)

- AMD research gift in support of research in the area of compilation for heterogeneous multicores (2011–)

- Gift from National Instruments in support of research on compiler optimizations for LabVIEW (2013–2015)

- INRIA Associate Team award (2012–2015) with Albert Cohen (INRIA / ENS)

- NVIDIA GPU research center award for 2012–2013, 2015-2016

- Research grants from Intel labs, Bengaluru (2013–2014) and from C-DAC, Bengaluru (2013–2014)

Students and Advising

- Ph.D.: 3 (1 best CS thesis medal), 2 (ongoing)

- M-Tech (Res.): 10 (4 best CS thesis medals)

- M-Tech: 8 (graduated)

Miscellaneous

- Program committee member: ASPLOS 2023-24, ASPLOS 2018, PLDI 2017 (distinguished reviewer award), Supercomputing 2016, PPoPP 2016, CC 2016, PPoPP 2012, IMPACT 2011–2016; External review committee: PPoPP 2024, PLDI 2014; Associate editor: ACM TACO

- Program chair: IMPACT 2012

- Reviewer: LCPC'06, PPoPP'07, ICS'07, LCPC'07, PACT'09, GPGPU workshop 2010, PPoPP 2011, HPCA 2011, IMPACT 2011–2016, ACM TOPLAS, ACM TACO, IEEE TPDS, JPDC

- Languages: English (fluent), Hindi (native), Telugu (native), French (intermediate), Kannada (intermediate).