

Matei Zaharia

Assistant Professor, Stanford Computer Science

Email: matei@cs.stanford.edu

Web: <https://cs.stanford.edu/~matei>

ACADEMIC HISTORY

- University of California at Berkeley (2007-2013)
Ph.D., Computer Science
Dissertation: [An Architecture for Fast and General Data Processing on Large Clusters](#)
Advisors: Scott Shenker and Ion Stoica
- University of Waterloo (2003-2007)
Bachelor of Mathematics, Honors Computer Science
Minor in Combinatorics and Optimization

EMPLOYMENT RECORD

- Assistant Professor of Computer Science, Stanford (2016-present)
- Assistant Professor of Computer Science, MIT (2015-2016)
- Chief Technologist and Cofounder, Databricks (2013-present)
- Graduate Research Assistant, University of California at Berkeley (2007-2013)

AWARDS AND HONORS

- Best Student Paper, Supercomputing 2021
- NSDI Test of Time Paper Award, 2021
- EuroSys Test of Time Paper Award, 2020
- Presidential Early Career Award for Scientists and Engineers (PECASE), 2019
- Best Student Paper, ICLR DebugML Workshop 2019
- Facebook Hardware & Software Systems Research Award, 2018
- NSF CAREER Award, 2017
- Best Student Paper, IEEE BigData 2017
- VMware Systems Research Award, 2016
- Best Paper Award Runner-Up, KDD 2016
- Google Research Award, 2015
- ACM Doctoral Dissertation Award, 2014
- U. Waterloo Faculty of Mathematics Young Alumni Achievement Medal, 2014
- Sort Benchmark Daytona GraySort World Record, 2014
- David J. Sakrison Prize for Research, UC Berkeley, 2013
- Best Paper Award, SIGCOMM 2012
- Best Paper Award and Honorable Mention for Community Award, NSDI 2012
- Best Demo Award, SIGMOD 2012
- Google Ph.D. Fellowship, 2011-2012
- NSERC Postgraduate Scholarship, 2009-2011
- Tong Leon Lim Pre-Doctoral Prize, UC Berkeley, 2009
- Julie Payette NSERC Research Scholarship, 2007-2009

PROFESSIONAL ACTIVITIES

- Technical Steering Committee Member, Linux Foundation MLflow Project (2020-present)
- Program Co-Chair, DISPA Workshop at VLDB 2020
- Program Co-Chair, MLOps Workshop at MLSys 2020
- Board Member, Conference on Systems and Machine Learning (2019-present)
- Technical Steering Committee Member, Linux Foundation Delta Lake Project (2019-present)
- Program Co-Chair, SysML 2019 Conference
- Co-Organizer, SysML 2018 Conference
- Vice President, Apache Spark Project (2013-present)
- Member, Association of Computing Machinery
- Member, Apache Software Foundation

PUBLICATIONS

Referred Journal Publications

1. Relevance-guided Supervision for OpenQA with ColBERT. O. Khattab, C. Potts and M. Zaharia. *TACL*, 2021.
2. Machine Learned Cellular Phenotypes in Cardiomyopathy Predict Sudden Death. A. Rogers, A. Selvalingam, M. Alhusseini, D. Krummen, C. Corrado, F. Abuzaid, T. Baykaner, C. Meyer, P. Clopton, W. Giles, P. Bailis, S. Niederer, P. Wang, W-J. Rappel, M. Zaharia and S. Narayan. *Circulation Research*, 128(2):172-184, January 2021.
3. DIFF: A Relational Interface for Large-Scale Data Explanation (extended version). F. Abuzaid, P. Kraft, S. Suri, E. Gan, E. Xu, A. Shenoy, A. Ananthanarayan, J. Sheu, E. Meijer, X. Wu, J. Naughton, P. Bailis, and M. Zaharia. *VLDB Journal*, 30:45-70, September 2020.
4. Machine Learning to Classify Intracardiac Electrical Patterns During Atrial Fibrillation. M. Alhusseini, F. Abuzaid, A. Rogers, J. Zaman, T. Baykaner, P. Clopton, P. Bailis, M. Zaharia, P. Wang, W-J. Rappel, and S. Narayan. *Circulation: Arrhythmia and Electrophysiology*, 13(8), August 2020.
5. Analysis of DAWNbench, a Time-to-Accuracy Machine Learning Performance Benchmark. C. Coleman, D. Kang, D. Narayanan, L. Nardi, T. Zhao, J. Zhang, P. Bailis, C. Olukotun, C. Re and M. Zaharia. *SIGOPS Operating Systems Review*, 53(1):14-25, July 2019.
6. Accelerating the Machine Learning Lifecycle with MLflow. M. Zaharia, A. Chen, A. Davidson, A. Ghodsi, S.A. Hong, A. Konwinski, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, F. Xie, and C. Zumar. *IEEE Data Engineering Bulletin*, 41(4), December 2018.
7. Apache Spark: A Unified Engine for Big Data Processing. M. Zaharia, R. Xin, P. Wendell, T. Das, M. Armbrust, A. Dave, X. Meng, J. Rosen, S. Venkataraman, M. Franklin, A. Ghodsi, J. Gonzalez, S. Shenker, I. Stoica. *Communications of the ACM*, 59(11):56-65, November 2016.
8. MLlib: Machine Learning in Apache Spark. X. Meng, J. Bradley, B. Yuvaz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D. Tsai, M. Amde, S. Owen, D. Xin, R. Xin, M. Franklin, R. Zadeh, M. Zaharia, and A. Talwalkar. *JMLR*, 17(34):1-7, April 2016.
9. A Cloud-Compatible Bioinformatics Pipeline for Ultrarapid Pathogen Identification from Next-Generation Sequencing of Clinical Samples. S.N. Naccache, S. Federman, N. Veeraraghavan, M. Zaharia, D. Lee, E. Samayoa, J. Bouquet, A.L. Greninger, K. Luk, B. Enge, D.A. Wadford, S.L. Messenger, G.L. Genrich, K. Pellegrino, G. Grard, E. Leroy, B.S. Schneider, J.N. Fair, M.A. Martinez, P. Isa, J.A. Crump, J.L. DeRisi, T. Sittler, J. Hackett Jr., S. Miller and C.Y. Chiu, *Genome Research*, 24(7):1180-92, June 2014.
10. Above the Clouds: A View of Cloud Computing. M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D.A. Patterson, A. Rabkin, I. Stoica and M. Zaharia, *Communications of the ACM*, 53(4):50-58, April 2010.

11. Design and Implementation of the KioskNet System. S. Guo, M. Derakhshani, M.H. Falaki, U. Ismail, R. Luk, E.A. Oliver, S. Ur Rahman, A. Seth, M.A. Zaharia, S. Keshav, *Computer Networks*, 55(1):264-281, January 2011.
12. Gossip-based Search Selection in Hybrid Peer-to-Peer Networks. M. Zaharia and S. Keshav, *J. Concurrency and Computation: Practice and Experience*, 20(2):139-153, February 2008.
13. Very Low-Cost Internet Access Using KioskNet. S. Guo, M.H. Falaki, E.A. Oliver, S. Ur Rahman, A. Seth, M. Zaharia, and S. Keshav, *ACM Computer Communication Review*, 37(5):95-100, October 2007.

Refereed Conference Publications

1. DBOS: A DBMS-Oriented Operating System. A. Skiadopoulou, Q. Li, P. Kraft, K. Kaffes, D. Hong, S. Matthew, D. Bestor, M. Cafarella, V. Gadepally, G. Graefe, J. Kepner, C. Kozyrakis, T. Kraska, M. Stonebraker, L. Suresh and M. Zaharia. VLDB 2022.
2. VIVA: An End-to-End System for Interactive Video Analytics. D. Kang, F. Romero, P. Bailis, C. Kozyrakis and M. Zaharia. CIDR 2022.
3. A Progress Report on DBOS: A Database-oriented Operating System. Q. Li, P. Kraft, K. Kaffes, A. Skiadopoulou, D. Kumar, J. Li, M. Cafarella, G. Graefe, J. Kepner, C. Kozyrakis, M. Stonebraker, L. Suresh and M. Zaharia. CIDR 2022.
4. Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval. O. Khattab, C. Potts and M. Zaharia. NeurIPS 2021 (Spotlight).
5. Solving Large-Scale Granular Resource Allocation Problems Efficiently with POP. D. Narayanan, F. Kazhmiaka, F. Abuzaid, P. Kraft, A. Agrawal, S. Kandula, S. Boyd and M. Zaharia. SOSP 2021.
6. Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM. D. Narayanan, M. Shoeybi, J. Casper, P. LeGresley, M. Patwary, V.A. Korthikanti, D. Vainbrand, P. Kashinkunti, J. Bernauer, B. Catanzaro, A. Phanishayee, M. Zaharia. Supercomputing 2021.
7. Clamor: Extending Functional Cluster Computing Frameworks with Fine-Grained Remote Memory Access. P. Thaker, H. Ayers, D. Raghavan, N. Niu, P. Levis and M. Zaharia. SoCC 2021.
8. Jointly Optimizing Preprocessing and Inference for DNN-based Visual Analytics. D. Kang, A. Mathur, T. Veeramacheneni, P. Bailis and M. Zaharia. VLDB 2021.
9. Accelerating Approximate Aggregation Queries with Expensive Predicates. D. Kang, J. Guibas, P. Bailis, T. Hashimoto, Y. Sun and M. Zaharia. VLDB 2021.
10. Memory-Efficient Pipeline-Parallel DNN Training. D. Narayanan, A. Phanishayee, K. Shi, X. Chen and M. Zaharia. ICML 2021.
11. Express: Lowering the Cost of Metadata-hiding Communication with Cryptographic Privacy. S. Eskandarian, H. Corrigan-Gibbs, M. Zaharia and D. Boneh. USENIX Security 2021.
12. Contracting Wide-area Network Topologies to Solve Flow Problems Quickly. F. Abuzaid, S. Kandula, B. Arzani, I. Menache, M. Zaharia. and P. Bailis NSDI 2021.
13. Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. M. Armbrust, A. Ghodsi, R. Xin and M. Zaharia. CIDR 2021.
14. Challenges and Opportunities for Autonomous Vehicle Query Systems. F. Kazhmiaka, M. Zaharia and P. Bailis. CIDR 2021.
15. FrugalML: How to Use ML Prediction APIs More Accurately and Cheaply. L. Chen, M. Zaharia and J. Zou. NeurIPS 2020 (Oral Presentation).
16. Heterogeneity-Aware Cluster Scheduling Policies for Deep Learning Workloads. D. Narayanan, K. Santhanam, F. Kazhmiaka, A. Phanishayee and M. Zaharia. OSDI 2020.
17. Sparse GPU Kernels for Deep Learning. T. Gale, M. Zaharia, C. Young and E. Elsen. Supercomputing 2020.
18. Delta Lake: High-Performance ACID Table Storage over Cloud Object Stores. M. Armbrust, T. Das, L. Sun, B. Yavuz, S. Zhu, M. Murthy, J. Torres, H. van Hovell, A. Ionescu, A. Luszczak, M. Switakowski,

- M. Szafranski, X. Li, T. Ueshin, M. Mokhtar, P. Boncz, A. Ghodsi, S. Paranjpye, P. Senster, R. Xin, M. Zaharia. VLDB 2020.
19. Approximate Selection with Guarantees using Proxies. D. Kang, E. Gan, P. Bailis, T. Hashimoto and M. Zaharia. VLDB 2020.
 20. BlazeIt: Optimizing Declarative Aggregation and Limit Queries for Neural Network-Based Video Analytics. D. Kang, P. Bailis and M. Zaharia. VLDB 2020.
 21. ObliDB: Oblivious Query Processing for Secure Databases. S. Eskandarian and M. Zaharia. VLDB 2020.
 22. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. O. Khattab and M. Zaharia. SIGIR 2020.
 23. POSH: A Data-Aware Shell. D. Raghavan, S. Fouladi, P. Levis and M. Zaharia. USENIX ATC 2020.
 24. Offload Annotations: Bringing Heterogeneous Computing to Existing Libraries and Workloads. G. Yuan, S. Palkar, D. Narayanan and M. Zaharia. USENIX ATC 2020.
 25. Spectral Lower Bounds on the I/O Complexity of Computation Graphs. S. Jain and M. Zaharia. SPAA 2020.
 26. Selection via Proxy: Efficient Data Selection for Deep Learning. C. Coleman, C. Yeh, S. Musmann, B. Mirzasoleiman, P. Bailis, P. Liang, J. Leskovec and M. Zaharia. ICLR 2020.
 27. Fleet: A Framework for Massively Parallel Streaming on FPGAs. J. Thomas, P. Hanrahan and M. Zaharia. ASPLOS 2020.
 28. Willump: A Statistically-Aware End-to-end Optimizer for Machine Learning Inference. P. Kraft, D. Kang, D. Narayanan, S. Palkar, P. Bailis and M. Zaharia. MLSys 2020.
 29. Model Assertions for Monitoring and Improving ML Models. D. Kang, D. Raghavan, P. Bailis and M. Zaharia. MLSys 2020.
 30. Improving the Accuracy, Scalability, and Performance of Graph Neural Networks with Roc. Z. Jia, S. Lin, M. Gao, M. Zaharia and A. Aiken. MLSys 2020.
 31. MLPerf Training Benchmark. P. Mattson, C. Cheng, C. Coleman, G. Diamos, P. Micikevicius, D. Patterson, H. Tang, G-Y. Wei, P. Bailis, V. Bittorf, D. Brooks, D. Chen, D. Dutta, U. Gupta, K. Hazelwood, A. Hock, X. Huang, B. Jia, D. Kang, D. Kanter, N. Kumar, J. Liao, D. Narayanan, T. Oguntebi, G. Pekhimenko, L. Pentecost, V. J. Reddi, T. Robie, T. St. John, C-J. Wu, L. Xu, C. Young, and M. Zaharia. MLSys 2020.
 32. Optimizing Data Intensive Computations in Existing Libraries with Split Annotations. S. Palkar and M. Zaharia. SOSP 2019.
 33. Generalized Pipeline Parallelism for DNN Training. D. Narayanan, A. Harlap, A. Phanishayee, V. Seshadri, N. Devanur, G. Ganger, P. Gibbons, and M. Zaharia. SOSP 2019.
 34. TASO: Optimizing Deep Learning Computation with Automated Generation of Graph Substitutions. Z. Jia, O. Padon, J. Thomas, T. Warszawski, M. Zaharia, and A. Aiken. SOSP 2019.
 35. DIFF: A Relational Interface for Large-Scale Data Explanation. F. Abuzaid, P. Kraft, S. Suri, E. Gan, E. Xu, A. Shenoy, A. Ananthanarayan, J. Sheu, E. Meijer, X. Wu, J. Naughton, P. Bailis, and M. Zaharia. VLDB 2019.
 36. From Laptop to Lambda: Outsourcing Everyday Jobs to Thousands of Transient Functional Containers. S. Fouladi, F. Romero, D. Iter, Q. Li, S. Chatterjee, C. Kozyrakis, M. Zaharia, and K. Winstein. USENIX ATC 2019.
 37. LIT: Learned Intermediate Representation Training for Model Compression. A. Koratana, D. Kang, P. Bailis and M. Zaharia. ICML 2019.
 38. To Index or Not to Index: Optimizing Exact Maximum Inner Product Search. F. Abuzaid, G. Sethi, P. Bailis and M. Zaharia. ICDE 2019.
 39. Beyond Data and Model Parallelism for Deep Neural Networks. Z. Jia, M. Zaharia and A. Aiken. SysML 2019.
 40. Optimizing DNN Computation with Relaxed Graph Substitutions. Z. Jia, J. Thomas, T. Warszawski, M. Gao, M. Zaharia and A. Aiken. SysML 2019.

41. Challenges and Opportunities in DNN-Based Video Analytics: A Demonstration of the BlazeIt Video Query Engine. D. Kang, P. Bailis and M. Zaharia. CIDR 2019.
42. Filter Before You Parse: Faster Analytics on Raw Data with Sparser. S. Palkar, F. Abuzaid, P. Bailis and M. Zaharia. VLDB 2018.
43. Evaluating End-to-End Optimization for Data Analytics Applications in Weld. S. Palkar, J. Thomas, D. Narayanan, P. Thaker, R. Palamuttam, P. Negi, A. Shanbhag, M. Schwarzkopf, H. Pirk, S. Amarasinghe, S. Madden and M. Zaharia. VLDB 2018.
44. MISTIQUE: A System to Store and Query Model Intermediates for Model Diagnosis. M. Vartak, J. da Trindade, S. Madden and M. Zaharia. SIGMOD 2018.
45. Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark. M. Armbrust, T. Das, J. Torres, B. Yavuz, S. Zhu, R. Xin, A. Ghodsi, I. Stoica and M. Zaharia. SIGMOD 2018.
46. Making Caches Work for Graph Analytics. Y. Zhang, V. Kiriansky, C. Mendis, M. Zaharia and S. Amarasinghe. IEEE BigData 2017.
47. DAWNbench: An End-to-End Deep Learning Benchmark and Competition. C. Coleman, D. Narayanan, D. Kang, T. Zhao, J. Zhang, L. Nardi, P. Bailis, K. Olukotun, C. Re, M. Zaharia. NIPS SysML 2017.
48. Stadium: A Distributed Metadata-Private Messaging System. N. Tyagi, Y. Gilad, D. Leung, M. Zaharia and N. Zeldovich. SOSP 2017.
49. NoScope: Optimizing Neural Network Queries over Video at Scale. D. Kang, J. Emmons, F. Abuzaid, P. Bailis and M. Zaharia. VLDB 2017.
50. Splinter: Practical Private Queries on Public Data. F. Wang, C. Yun, S. Goldwasser, V. Vaikuntanathan and M. Zaharia. NSDI 2017.
51. Weld: A Common Runtime for High Performance Data Analytics. S. Palkar, J. Thomas, A. Shanbhag, D. Narayanan, H. Pirk, M. Schwarzkopf, S. Amarasinghe and M. Zaharia. CIDR 2017.
52. Yggdrasil: An Optimized System for Training Deep Decision Trees at Scale. F. Abuzaid, J. Bradley, F. Liang, A. Feng, L. Yang, M. Zaharia and A. Talwalkar. NIPS 2016.
53. Voodoo – A Vector Algebra for Portable Database Performance on Modern Hardware. H. Pirk, O. Moll, M. Zaharia and S. Madden. VLDB 2016.
54. Matrix Computations and Optimizations in Apache Spark. R.B. Zadeh, X. Meng, A. Staple, B. Yavuz, L. Pu, S. Venkataraman, E. Sparks, A. Ulanov and M. Zaharia. KDD 2016.
55. SparkR: Scaling R Programs with Spark. S. Venkataraman, Z. Yang, D. Liu, E. Liang, X. Meng, R. Xin, A. Ghodsi, M. Franklin, I. Stoica and M. Zaharia. SIGMOD 2016.
56. FairRide: Near-Optimal, Fair Cache Sharing. Q. Pu, H. Li, M. Zaharia, A. Ghodsi, and I. Stoica. NSDI 2016.
57. Vuvuzela: Scalable Private Messaging Resistant to Traffic Analysis. J. van den Hooff, D. Lazar, M. Zaharia and N. Zeldovich. SOSP 2015, October 2015.
58. Scaling Spark in the Real World: Performance and Usability. M. Armbrust, T. Das, A. Davidson, A. Ghodsi, A. Or, J. Rosen, I. Stoica, P. Wendell, R. Xin and M. Zaharia. VLDB 2015, August 2015.
59. Spark SQL: Relational Data Processing in Spark. M. Armbrust, R. Xin, C. Lian, Y. Huai, D. Liu, J. Bradley, X. Meng, T. Kaftan, M. Franklin, A. Ghodsi and M. Zaharia. SIGMOD 2015, June 2015.
60. Tachyon: Reliable, Memory Speed Storage for Cluster Computing Frameworks. H. Li, A. Ghodsi, M. Zaharia, S. Shenker and I. Stoica, SOCC 2014, November 2014.
61. Discretized Streams: Fault-Tolerant Streaming Computation at Scale. M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica. SOSP 2013, November 2013.
62. Sparrow: Distributed, Low-Latency Scheduling. K. Ousterhout, P. Wendell, M. Zaharia and I. Stoica. SOSP 2013, November 2013.
63. Shark: SQL and Rich Analytics at Scale. R. Xin, J. Rosen, M. Zaharia, M. Franklin, S. Shenker, and I. Stoica. SIGMOD 2013, June 2013.
64. Choosy: Max-Min Fair Sharing for Datacenter Jobs with Constraints. A. Ghodsi, M. Zaharia, S. Shenker and I. Stoica. EuroSys 2013, April 2013.

65. Multi-Resource Fair Queueing for Packet Processing. A. Ghodsi, V. Sekar, M. Zaharia and I. Stoica. SIGCOMM 2012, August 2012.
66. Optimally Designing Games for Cognitive Science Research. A.N. Rafferty, M. Zaharia and T.L. Griffiths. Annual Conf. of the Cognitive Science Society, August 2012.
67. Cloud Terminal: Secure Access to Sensitive Applications from Untrusted Systems. L. Martignoni, P. Poosankam, M. Zaharia, J. Han, S. McCamant, D. Song, V. Paxson, A. Perrig, S. Shenker, I. Stoica. USENIX ATC 2012, June 2012.
68. Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica. NSDI 2012, April 2012.
69. Scaling the Mobile Millennium System in the Cloud. T. Hunter, T. Moldovan, M. Zaharia, S. Merzgui, J. Ma, M.J. Franklin, P. Abbeel, and A.M. Bayen. SOCC 2011, October 2011.
70. Managing Data Transfers in Computer Clusters with Orchestra. M. Chowdhury, M. Zaharia, J. Ma, M.I. Jordan and I. Stoica, SIGCOMM 2011, August 2011.
71. Mesos: A Platform for Fine-Grained Resource Sharing in the Data Center. B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A.D. Joseph, R. Katz, S. Shenker and I. Stoica, NSDI 2011, March 2011.
72. Dominant Resource Fairness: Fair Allocation of Multiple Resources Types. A. Ghodsi, M. Zaharia, B. Hindman, A. Konwinski, S. Shenker, and I. Stoica, NSDI 2011, March 2011.
73. Delay Scheduling: A Simple Technique for Achieving Locality and Fairness in Cluster Scheduling. M. Zaharia, D. Borthakur, J. Sen Sarma, K. Elmeleegy, S. Shenker and I. Stoica. EuroSys 2010, April 2010.
74. ICTD for Healthcare in Ghana: Two Parallel Case Studies. R. Luk, M. Zaharia, M. Ho, B. Levine and P. Aoki, ICTD 2009, April 2009.
75. Improving MapReduce Performance in Heterogeneous Environments. M. Zaharia, A. Konwinski, A.D. Joseph, R. Katz and I. Stoica, OSDI 2008, December 2008.
76. Design and Implementation of the KioskNet System. S. Guo, M.H. Falaki, E.A. Oliver, S. Ur Rahman, A. Seth, M. Zaharia, U. Ismail, and S. Keshav, ICTD 2007, December 2007.

Refereed Workshop Publications

1. What can Data-Centric AI Learn from Data and ML Engineering?. N. Polyzotis and M. Zaharia. NeurIPS Data-Centric AI Workshop 2021.
2. Finding Label Errors in Autonomous Vehicle Data With Learned Observation Assertions. D. Kang, N. Arechiga, S. Pillai, P. Bailis and M. Zaharia. NeurIPS Data-Centric AI Workshop 2021.
3. Exploiting Proximity Search and Easy Examples to Select Rare Events. D. Kang, A. Derhacopian, K. Tsuji, T. Hebert, P. Bailis, T. Fukami, T. Hashimoto, Y. Sun and M. Zaharia. NeurIPS Data-Centric AI Workshop 2021.
4. Don't Hate the Player, Hate the Game: Safety and Utility in Multi-Agent Congestion Control. P. Thaker, M. Zaharia and T. Hashimoto. HotNets 2021.
5. Finding Label and Model Errors in Perception Data with Learned Observation Assertions (Extended Abstract). D. Kang, N. Arechiga, S. Pillai, P. Bailis and M. Zaharia. VLDB AIDB Workshop 2021.
6. Breakfast of Champions: Towards Zero-Copy Serialization with NIC Scatter-Gather. D. Raghavan, P. Levis, M. Zaharia and I. Zhang. HotOS 2021.
7. A Polystore Based Database Operating System. M. Cafarella, D. DeWitt, V. Gadepally, J. Kepner, C. Kozyrakis, T. Kraska, M. Stonebraker and Matei Zaharia. VLDB Poly Workshop 2020.
8. Analysis and Exploitation of Dynamic Pricing in the Public Cloud for ML Training. D. Narayanan, K. Santhanam, F. Kazhamiaka, A. Phanishayee and M. Zaharia. VLDB DISPA Workshop 2020.
9. To Call or not to Call? Using ML Prediction APIs more Accurately and Economically. L. Chen, M. Zaharia and J. Zou. ICML EcoPaDL Workshop 2020.

10. Developments in MLflow: A System to Accelerate the Machine Learning Lifecycle. A. Chen, A. Chow, A. Davidson, A. DCuncha, A. Ghodsi, S.A. Hong, A. Konwinski, C. Mewald, S. Murching, T. Nykodym, P. Ogilvie, M. Parkhe, A. Singh, F. Xie, M. Zaharia, R. Zang, J. Zheng and C. Zumar. SIGMOD DEEM Workshop 2020.
11. Debugging Machine Learning via Model Assertions. D. Kang, D. Raghavan, P. Bailis and M. Zaharia. ICLR DebugML Workshop 2019.
12. Model Assertions for Debugging Machine Learning. D. Kang, D. Raghavan, P. Bailis and M. Zaharia. NeurIPS Systems for ML Workshop 2018.
13. Analysis of the Time-To-Accuracy Metric and Entries in the DAWNbench Deep Learning Benchmark. C. Coleman, D. Kang, D. Narayanan, L. Nardi, T. Zhao, J. Zhang, P. Bailis, K. Olukotun, C. Re and M. Zaharia. NeurIPS Systems for ML Workshop 2018.
14. Accelerating Deep Learning Workloads through Efficient Multi-Model Execution. D. Narayanan, K. Santhanam, A. Phanishayee and M. Zaharia. NeurIPS Systems for ML Workshop 2018.
15. Exploring the Use of Learning Algorithms for Efficient Performance Profiling. S. Palkar, S. Suri, P. Bailis and M. Zaharia. NeurIPS ML for Systems Workshop 2018.
16. Block-wise Intermediate Representation Training for Model Compression. A. Koratana, D. Kang, P. Bailis and M. Zaharia. NeurIPS CDNNRIA Workshop 2018.
17. DIY Hosting for Online Privacy. S. Palkar and M. Zaharia. HotNets 2017.
18. GraphFrames: An Integrated API for Mixing Graph and Relational Queries. A. Dave, A. Jindal, L.E. Li, R. Xin, J. Gonzalez and M. Zaharia. GRADES 2016.
19. ModelDB: A System for Machine Learning Model Management. M. Vartak, H. Subramanyam, W.E. Lee, S. Viswanathan, S. Husnoo, S. Madden and M. Zaharia. HILDA 2016.
20. Discretized Streams: An Efficient and Fault-Tolerant Model for Stream Processing on Large Clusters. M. Zaharia, T. Das, H. Li, S. Shenker and I. Stoica. HotCloud 2012, June 2012.
21. The Datacenter Needs an Operating System. M. Zaharia, B. Hindman, A. Konwinski, A. Ghodsi, A.D. Joseph, R. Katz, S. Shenker and I. Stoica, HotCloud 2011, June 2011.
22. Spark: Cluster Computing with Working Sets. M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica. HotCloud 2010, June 2010.
23. A Common Substrate for Cluster Computing. B. Hindman, A. Konwinski, M. Zaharia and I. Stoica, HotCloud 2009, June 2009.
24. Finding Content in File-Sharing Networks When You Can't Even Spell. M. Zaharia, A. Chandel, S. Saroiu, and S. Keshav, Proc. IPTPS, February 2007.

Non-refereed Publications

1. A Moderate Proposal for Radically Better AI-Powered Web Search. O. Kahattab, C. Potts and M. Zaharia. Stanford HAI Blog, 2021.
2. POSH: A Data-Aware Shell. D. Raghavan, S. Fouladi, P. Levis, and M. Zaharia. USENIX ;login:, 45(4), December 2020.
3. Outsourcing Everyday Jobs to Thousands of Cloud Functions with gg. S. Fouladi, F. Romero, D. Iter, Q. Li, S. Chatterjee, C. Kozyrakis, M. Zaharia, and K. Winstein. USENIX ;login:, 44(3), September 2019.
4. Fast and Interactive Analytics over Hadoop Data with Spark. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M.J. Franklin, S. Shenker, I. Stoica. USENIX ;login:, August 2012.
5. Mesos: Flexible Resource Sharing for the Cloud. B. Hindman, A. Konwinski, M. Zaharia, A. Ghodsi, A.D. Joseph, R. Katz, S. Shenker and I. Stoica, USENIX ;login:, August 2011.

Books in Print

1. Spark: The Definitive Guide: Big Data Processing Made Simple. B. Chambers, M. Zaharia. O'Reilly Media, 2018.
2. Learning Spark: Lightning-Fast Big Data Analysis. H. Karau, A. Konwinski, P. Wendell, M. Zaharia. O'Reilly Media, 2015.
3. An Architecture for Fast and General Data Processing on Large Clusters. M. Zaharia. ACM and Morgan & Claypool, 2016.

PATENTS

1. Structured cluster execution for data streams. M. Armbrust, T. Das, S. Xin, M. Zaharia. US Patent 10558664, issued 2020.
2. Serverless execution of code using cluster resources. A. Ghodsi, S. Shankar, S. Paranjpye, S. Xin, M. Zaharia. US patent 10474501, issued 2019.
3. Cluster instance management system. A. Ghodsi, I. Stoica, M. Zaharia. US Patent 9769032, issued 2017.
4. System for exploring data in a database. A. Ghodsi, I. Stoica, M. Zaharia. US Patent 10095735, issued 2017.

STUDENTS

Current Ph.D. Students

1. Lingjiao Chen (expected graduation: 2026)
2. Omar Khattab (expected graduation: 2026)
3. Keshav Santhanam (expected graduation: 2026)
4. Trevor Gale (expected graduation: 2025)
5. Gina Yuan (expected graduation: 2025)
6. Deepti Raghavan (expected graduation: 2024)
7. Peter Kraft (expected graduation: 2024)
8. Firas Abuzaid (expected graduation: 2022)
9. Daniel Kang (expected graduation: 2022)
10. Pratiksha Thaker (expected graduation: 2022)
11. James Thomas (expected graduation: 2022)

Former Ph.D. Students

1. Deepak Narayanan (graduated 2021, now at Microsoft Research)
Dissertation: Resource-Efficient Execution of Deep Learning Computations
2. Cody Coleman (graduated 2021, now at Coactive)
Dissertation: Resource and Data Efficient Deep Learning
3. Zhihao Jia (graduated 2020, now at CMU)
Dissertation: Automated Discovery of Machine Learning Optimizations
4. Shoumik Palkar (graduated 2020, now at Databricks)
Dissertation: Interfaces for Efficient Software Composition on Modern Hardware

Current Postdocs

1. Fiodar Kazhamiaka (appointed: 2020)

Undergraduate and M.S. Students with Refereed Publications

1. Animesh Koratana (undergrad, expected graduation: 2022)
2. John Guibas (undergrad, expected graduation: 2022)
3. Christopher Yeh (M.S., graduated 2019)
4. Ankit Mathur (M.S., graduated 2019)
5. Saachi Jain (M.S., graduated 2019)
6. Keshav Santhanam (M.S., graduated 2019)
7. Parimajan Negi (undergrad, graduated 2018)
8. Rahul Palamuttam (M.S., graduated 2018)
9. Catherine Yun (M.S. at MIT, graduated 2017)
10. Nirvan Tyagi (M.S. at MIT, graduated 2016)