Deepak Narayanan

Curriculum Vitae

✉ *deepakn@stanford.edu*
🖱 *https://cs.stanford.edu/∼deepakn/*

Education

2015-Present **Ph.D., Stanford University**, Computer Science
Advisor: Matei Zaharia
Thesis title: Resource-Efficient Execution of Deep Learning Computations
Research areas: Systems for Machine Learning, Distributed Systems, Cloud Computing, Performance Optimization.
I broadly work on Systems. I design and implement software to improve the *runtime performance* and *runtime effeciency* of emerging machine learning and data analytics workloads on modern hardware.

2014-2015 **Master of Engineering, Massachusetts Institute of Technology**, GPA: 5.0/5.0, Computer Science.

2011-2013 **Bachelor of Science, Massachusetts Institute of Technology**, GPA: 4.9/5.0, Computer Science and Mathematics.

Selected Honors

2016-2021 NSF Graduate Research Fellowship.

2015-2016 Stanford Graduate Fellowship.

2011 All India Rank 229 in IIT Joint Entrance Examination (∼500,000 candidates).

2010-2011 International Mathematics Olympiad Training Camp Invitee, Government of India (~ 50 invitees across India).

2010 Kishore Vaigyanik Protsahan Yojana (KVPY) Fellowship, Government of India ($\sim 250/100,000$ candidates across India).

2009 International Astronomy Olympiad (Junior) Training Camp Invitee, Government of India (~ 30 invitees across India).

2007 National Talent Search Examination (NTSE) Scholarship, Government of India (~ 1000 scholars across India).

2007 Study of Exceptional Talent (SET) Membership, Center for Talented Youth (CTY), Johns Hopkins University.

Publications

Conferences **Heterogeneity-Aware Cluster Scheduling Policies for Deep Learning Workloads**
Deepak Narayanan, Keshav Santhanam, Fiodar Kazhamiaka, Amar Phanishayee, Matei Zaharia. *OSDI 2020*.

Offload Annotations: Bringing Heterogeneous Computing to Existing Libraries and Workloads
Gina Yuan, Shoumik Palkar, **Deepak Narayanan**, Matei Zaharia. *USENIX ATC 2020*.

Willump: A Statistically-Aware End-to-end Optimizer for Machine Learning Inference
Peter Kraft, Daniel Kang, **Deepak Narayanan**, Shoumik Palkar, Peter Bailis, Matei Zaharia. *MLSys 2020*.

MLPerf Training Benchmark
Peter Mattson, Christine Cheng, Cody Coleman, Greg Diamos, Paulius Micikevicius, David Patterson, Hanlin Tang, Gu-Yeon Wei, Peter Bailis, Victor Bittorf, David Brooks, Dehao Chen, Debojyoti Dutta, Udit Gupta, Kim Hazelwood, Andrew Hock, Xinyuan Huang, Bill Jia, Daniel Kang, David Kanter, Naveen Kumar, Jeffery Liao, **Deepak Narayanan**, Tayo Oguntebi, Gennady Pekhimenko, Lillian Pentecost, Vijay Janapa Reddi, Taylor Robie, Tom St. John, Carole-Jean Wu, Lingjie Xu, Cliff Young, Matei Zaharia. *MLSys 2020*.

PipeDream: Generalized Pipeline Parallelism for DNN Training
Deepak Narayanan*, Aaron Harlap*, Amar Phanishayee, Vivek Seshadri, Nikhil R. Devanur, Gregory R. Ganger, Phillip B. Gibbons, Matei Zaharia. *SOSP 2019*.

Evaluating End-to-End Optimization for Data Analytics Applications in Weld
Shoumik Palkar, James Thomas, **Deepak Narayanan**, Pratiksha Thaker, Parimarjan Negi, Rahul Palamuttam, Anil Shanbhag, Holger Pirk, Malte Schwarzkopf, Saman Amarasinghe, Samuel Madden, Matei Zaharia. *VLDB 2018*.

MacroBase: Prioritizing Attention in Fast Data
Peter Bailis, Edward Gan, Samuel Madden, **Deepak Narayanan**, Kexin Rong, Sahaana Suri. *SIGMOD 2017*.

Weld: A Common Runtime for High Performance Data Analytics
Shoumik Palkar, James Thomas, Anil Shanbhag, **Deepak Narayanan**, Holger Pirk, Malte Schwarzkopf, Saman Amarasinghe, Matei Zaharia. *CIDR 2017*.

Journals

Analysis of DAWNBench, a Time-to-Accuracy Machine Learning Performance Benchmark
Cody Coleman*, Daniel Kang*, **Deepak Narayanan***, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Re, Matei Zaharia. *SIGOPS Operating Systems Review July 2019*.

MacroBase: Prioritizing Attention in Fast Data
Firas Abuzaid, Peter Bailis, Jialin Ding, Edward Gan, Samuel Madden, **Deepak Narayanan**, Kexin Rong, Sahaana Suri. *TODS 2018*.

Workshops

Analysis and Exploitation of Dynamic Pricing in the Public Cloud for ML Training
Deepak Narayanan, Keshav Santhanam, Fiodar Kazhamiaka, Amar Phanishayee, Matei Zaharia. *DISPA 2020*.

Accelerating Deep Learning Workloads through Efficient Multi-Model Execution
Deepak Narayanan, Keshav Santhanam, Amar Phanishayee, Matei Zaharia. *NeurIPS Systems for ML Workshop 2018*.

Analysis of the Time-To-Accuracy Metric and Entries in the DAWNBench Deep Learning Benchmark
Cody Coleman*, Daniel Kang*, **Deepak Narayanan***, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Re, Matei Zaharia. *NeurIPS Systems for ML Workshop 2018*.

DAWNBench: An End-to-End Deep Learning Benchmark and Competition
Cody Coleman, **Deepak Narayanan**, Daniel Kang, Tian Zhao, Jian Zhang, Luigi Nardi, Peter Bailis, Kunle Olukotun, Christopher Re, Matei Zaharia. *NeurIPS Systems for ML Workshop 2017*.

Preprints **Memory-Efficient Pipeline-Parallel DNN Training**
Deepak Narayanan, Amar Phanishayee, Kaiyu Shi, Xie Chen, Matei Zaharia. *arXiv:2006.09503*.

Weld: Rethinking the Interface Between Data-Intensive Libraries
Shoumik Palkar, James Thomas, **Deepak Narayanan**, Anil Shanbhag, Holger Pirk, Malte Schwarzkopf, Saman Amarasinghe, Samuel Madden, Matei Zaharia. *arXiv:1709.06416*.

Teaching

CS149 **Parallel Computing**, Stanford University, Fall 2019. Instructor: Kayvon Fatahalian, Kunle Olukotun.

CS245 **Principles of Data-Intensive Systems**, Stanford University, Spring 2019. Instructor: Matei Zaharia.

CS161 **Design and Analysis of Algorithms**, Stanford University, Fall 2018. Instructor: Aviad Rubinstein.

6.046 **Design and Analysis of Algorithms**, Massachusetts Institute of Technology, Spring 2015, Instructors: Erik Demaine, Srini Devadas, Nancy Lynch.

6.006 **Introduction to Algorithms**, Massachusetts Institute of Technology, Spring 2014, Instructors: Srini Devadas, Nancy Lynch, Vinod Vaikuntanathan.

Experience

Summer-Fall 2020 **Research Intern**, *NVIDIA*, Stanford, CA, Applied Deep Learning.
Mentor: Patrick LeGresley and Mohammad Shoeybi
Integrated pipeline parallelism primitives from PipeDream with intra-stage model parallelism used in Megatron.

Summer 2019 **Research Intern**, *Microsoft Research*, Redmond, WA, Systems and Networking.
Mentor: Amar Phanishayee
Explored applying pipeline parallelism to large models that do not fit on a single worker. Also worked on heterogeneous cluster scheduling.

Summer 2018 **Research Intern**, *Microsoft Research*, Redmond, WA, Systems and Networking.
Mentor: Amar Phanishayee
Generalized pipeline parallelism to modern model architectures and hardware topologies with heterogeneous communication links. Also helped build a performance debugging tool for deep learning applications.

Summer 2016 **Research Intern**, *Microsoft Research*, Redmond, WA, Systems and Networking.
Mentor: Amar Phanishayee
Explored the implications of combining model and data parallelism with input pipelining for deep learning training (PipeDream).

Summer 2015 **Research Intern**, *Microsoft Research*, Redmond, WA, Systems and Networking.
Mentor: Amar Phanishayee
Explored the effects of different synchronization primitives on deep model training on a multicore server.

Summer 2014 **Software Engineering Intern**, *Microsoft*, Bellevue, WA, Bing Ads Click Prediction.
Implemented a Machine Learning experimentation framework that allowed engineers to experiment with different click prediction indicators more efficiently.

Summer 2013 **Software Engineering Intern**, *Pinterest*, San Francisco, CA, Data Infrastructure.
Improved the reliability and efficiency of the company's post-hoc data analysis workflows.

Summer 2012 **Undergraduate Research Intern**, *Microsoft Research*, Bangalore, India, Multilingual Systems.
Mentor: Raghavendra Udupa
Designed and implemented a type-ahead email search system in C# for Microsoft Outlook that used NLP techniques to return context-sensitive, personalized suggestions to the user.

Talks

Nov. 2020 "PipeDream: Generalized Pipeline Parallelism for DNN Training".
Data+AI Summit Europe. Online (due to COVID-19).

Nov. 2020 "Resource-Efficient Execution of Deep Learning Computations".
Stanford SystemX. Online (due to COVID-19).

Jun. 2020 "Memory-Efficient Pipeline-Parallel DNN Training using PipeDream-2BW".
NVIDIA. Online (due to COVID-19).

Jun. 2020 "Memory-Efficient Pipeline-Parallel DNN Training using PipeDream-2BW".
Facebook. Online (due to COVID-19).

Jun. 2020 "Heterogeneity-Aware Cluster Scheduling Policies for Deep Learning Workloads.".
DAWN Retreat. Online (due to COVID-19).

Nov. 2019 "PipeDream: Generalized Pipeline Parallelism for DNN Training".
Facebook. Menlo Park, CA.

Sep. 2019 "PipeDream: Generalized Pipeline Parallelism for DNN Training".
DAWN Retreat. Menlo Park, CA.

Mar. 2019 "A Scheduler for Efficiently Sharing GPU Clusters".
DAWN Retreat. Chaminade, CA

Mar. 2018 "Accelerating Model Search with Model Batching".
DAWN Retreat. Santa Cruz, CA.

Nov. 2017 "Kostos: A Cost-Based Optimizer for Modern Hardware".
Stanford SystemX. Stanford, CA.

Sep. 2017 "Kostos: Cost-Based Optimization of Data Science Workloads".
DAWN Retreat. Santa Cruz, CA.

Mar. 2017 "Weld: A Common Runtime for Data Analytics".
Strata + Hadoop World. San Jose, CA.