Ameer Haj-Ali

Email: hajali.ameer@gmail.com
Personal Website: ameerhajali.com
Linkedin: ameer-haj-ali
Phone: +1 (971) 336-2824
Visa: **O1+EB1**

Education

University of California, Berkeley, California, USA
Ph.D. Student, Electrical Engineering and Computer Science, 2018-2020
Field: Compiler Optimization with Machine Learning
- **Finished five-year track *summa cum laude* in two years.**

Thesis Advisors: Prof. Krste Asanović and Prof. Ion Stoica
Minor: Business Administration

Israel Institute of Technology, Technion, Haifa, Israel
M.Sc., Electrical Engineering, Excellence Direct Track, June 2018, **The Valedictorian**
Thesis Title: Performing Image Processing in Memristive Memory
Thesis Advisor: Prof. Shahar Kvatinsky
- **Finished two-year track *summa cum laude* (top 3%) in one year.**
- **Nominated for Cadence Academic Network Master Thesis Award.**

B.Sc., Computer Engineering, May 2017, **The Valedictorian**
- **Finished four-year track *summa cum laude* (top 3%) in three years.**
- **Member of the President's List of highest honors for excellent scholastic achievements every semester**.

Professional Experience

Anyscale, Inc, San Francisco, California, USA.
L4 SWE → → **L5 Sr. SWE** → **TL** → **TLM** → **EM 1** → **EM 2**,→ **Sr. EM** 2019-present
"The fastest execution manager at Anyscale", "In execution, among the best I know".
Promoted from L4 IC to L7 Sr. EM in 3.33 years.

L6 – EM II, 12/01/2021-01/01/2024, **L7 – Sr. EM** 01/01/2024-present

- Led Anyscale Gen AI, LLM performance, and Endpoints teams (demo).
- Led Anyscale Gen AI infrastructure and serving teams.
- Anyscale growth eng team lead.
- Grew my team from 0 to 25 in less than 1 year. Managing through managers.
- The objective owner and project manager of Anyscale General Availability (GA) (15+ projects). Responsible for the sales, marketing, engineering, and product management departments to deliver the Anyscale GA product (40 Eng/TL/EM, 4 PMs, 3 SAs, 2 Security, 4 marketing/sales. Ex. Google/Uber/Microsoft/Facebook/LinkedIn/Amazon/etc). My demo video of the product.
- Leading the cloud platform engineering (2 Tech Lead/Architect, 2 PMs, 2 Managers, 18 L3-L6 SWEs), which builds the foundational blocks of Anyscale's serverless infrastructure end-to-end. This includes cluster orchestration, autoscaling, logging, metrics, billing, and a multi-cloud, multi-region architecture that provides a reliable and scalable managed Ray experience for Anyscale customers.
- Project manager of Anyscale's All-In-Our-Account (AIOA) and All-In-Customer-Account (AICA) projects.
  AIOA is a VM-based infrastructure completely managed by Anyscale (Anyscale Cloud) that customers can run on.
  AICA is a VM-based infrastructure managed by Anyscale but runs in the customer's account.

L5 Tech Lead Manager, 03/03/2021-12/01/2021
- Led Tiger Team 1, Tiger Team 2, and the most influential projects at Anyscale (by that time). I was overseeing 25+ Engineers, 7 Managers, 5 PMs, SREs, and 3 Product Designers. These projects productionized multiple new Anyscale products "in the fastest execution seen at Anyscale". This includes a new UI, frontend, backend, CLI, compute environments, data pipeline, and transitioning from open source Ray to the product with zero code changes. A demo video of the final product was broadcasted in the Ray Summit 2021.
- Led the Serverless team (7 Eng, 1 PM, ex. Google/Uber/MS/FB/Amazon/etc), which develops and maintains the Ray Cluster and Autoscaler, Ray Client, Cloud providers, and Ray on Kubernetes. The team develops the core engine of Anyscale's product for providing the infinite laptop serverless experience, used by every user of Ray and the proprietary product.
- Led the development of multiple features that reduced the company's cloud providers' bill by more than $0.5M/year (growing linearly with the number of employees).
- My team works closely with customers to address all their needs and concerns as fast as possible.

L5 Tech Lead, 12/11/2020-03/03/2021
- Built the new autoscaling infrastructure in Anyscale's product.
- Led the development of the open source Ray K8s operator.
- Led the development of C++ API for Ray to allow Anyscale's users to run distributed C++ applications. This is being used by multiple companies including Intel, Ant Financial, and ByteDance.

SWE L4 → L5 Sr. SWE, 11/15/2019-12/11/2020 (part time → full time on August 2020)
- Led multiple projects requested by Anyscale's early customers to get them on board.
- Built a cluster management system to allow customers to run Ray on their on premise clusters.
- Wrote the latest resource demand based open source Ray Autoscaler (video) used by most Ray users to scale up from a single node to multi-node.
- Contributed to RLlib, the state-of-the-art open-source library for running reinforcement learning.
- Implemented distributed Scikit-learn on top of Ray that runs on large clusters.
- Directed the first Ray meetup in Israel.

Anyscale is an early stage startup with 80 employees. The company raised $60.6M in one year. The company's goal is to make distributed computing accessible to everyone by building a commercial product out of Ray. Ray is one of the fastest growing open-source projects created by the founders of Anyscale with millions of users and more than 18,000 starts on GitHub and more than 100 contributors, including Uber, Microsoft, Amazon, Ant Financial, and Intel.

Intel Labs, Hillsboro, Oregon, USA
Artificial Intelligence Research Intern, Summer 2019
I started and lead an initiative for collaboration between research and engineering teams. We built two frameworks that use artificial intelligence to automatically optimize the performance of workloads for thousands of Intel users on personal laptops and in the cloud. This resulted in two successful projects:
- NeuroVectorizer: End-to-End Vectorization with Deep RL. **Published in top conference on compilers (CGO 2020), CATC 2019, and Workshop on ML for Systems at NeurIPS 2019.**
- RLDRM: Closed Loop Dynamic Cache Allocation with Deep Reinforcement Learning for Network Function Virtualization. **Published in NetSoft 2020. Best paper award**.

Nvidia (Previously Mellanox Technologies), Yokneam, Israel
Chip Designer, R&D, 2015-2016
Mellanox was a supplier of computer networking products based on InfiniBand and Ethernet technology. It had 2500 employees and $1B annual revenue.
- I worked on creating design and automation tools that automated the chip verification process.
- I lead the automation of integrating Ezchip and Mellanox chip designs after their merger.

American Technion Society (ATS), San Fransisco, California
Board Member, 2019-Present
Based in New York City, ATS provides critical support to the Technion - Israel Institute of Technology. ATS donors have provided more than $1.95 billion since its inception in 1940. I work on helping ATS fund raise by hosting events for potential donors and directing strategic decisions on the board.

Research

University of California, Berkeley, California, USA
Ph.D. student, working with Prof. Krste Asanović and Prof. Ion Stoica
Research Interests: Machine Learning, Reinforcement Learning, Code Optimization Hardware/Software Codesign, Auto-Tuning, and High Performance Computing
- Led/co-lead many projects spanning machine learning in compiler optimization and hardware software codesign. This includes Gemmini (funded by DARPA), AutoPhase, NeuroVectorizer, ProTuner, Ansor, AutoCkt and more (checkout my publications).
- Among the first researchers to introduce and build open source systems (e.g., AutoPhase, NeuroVectorizer, ProTuner) that use reinforcement learning in compiler optimization. Part of this research is being used by startups to automatically speedup the performance of systems. NeuroVectorizer was adopted and open sourced by official Intel Github.

Research

University of California, Berkeley, California, USA
Ph.D. student, working with Prof. Krste Asanović and Prof. Ion Stoica
Research Interests: Machine Learning, Reinforcement Learning, Code Optimization Hardware/Software Codesign, Auto-Tuning, and High Performance Computing
- Led/co-lead many projects spanning machine learning in compiler optimization and hardware software codesign. This includes Gemmini (funded by DARPA), AutoPhase, NeuroVectorizer, ProTuner, Ansor, AutoCkt and more (checkout my publications).
- Among the first researchers to introduce and build open source systems (e.g., AutoPhase, NeuroVectorizer, ProTuner) that use reinforcement learning in compiler optimization. Part of this research is being used by startups to automatically speedup the performance of systems. NeuroVectorizer was adopted and open sourced by official Intel Github.

Israel Institute of Technology, Technion, Haifa, Israel
Graduate student working with Prof. Shahar Kvatinsky, 2016-2018
Project: Memristive Memory Processing Unit (mMPU)
- Design of high performance, energy efficient processing in-memory system using emerging memory technologies.
  This work has been cited more than 200 times.

Undergraduate research student with Prof. Shahar Kvatinsky, 2016
Project 1: Configurable, Multi-threaded, Cycle Accurate, Processor Simulator
- Collaborative work with researchers from CEVA, Inc., to enhance the performance of CEVA's digital signal processors.

Project 2: GPU Algorithms Development for Image Processing

Awards & Fellowships

1. Won first place in Anyscale Hackathon 2020.
2. The Valedictorian Honor (M.Sc.), Technion, 2019.
3. Open Gateway Fellowship, UC Berkeley, 2018.
4. The William Oldham Fellowship, UC Berkeley, 2018.
5. The Valedictorian Honor (B.Sc.), Technion, 2017.
6. Dean's scholarship for excellent graduate students, Technion, 2016.
7. Full tuition scholarship for M.Sc. studies, Technion , 2016-2018.
8. The System Architecture Labs Cluster Prize for outstanding undergraduate projects (received twice), Technion, 2016.
9. Excellence award from Apple for excellent scholastic achievements, Technion, 2016.
10. Member of the President's List of highest honors for excellent scholastic achievements in all undergraduate semesters (top 3%), Technion, 2013-2016.
11. Full tuition scholarship for B.Sc. studies, Technion, 2013-2016.

Teaching

Electrical Engineering and Computer Science, University of California, Berkeley
1. Graduate Teaching Assistance, Introduction to Machine Learning, Spring 2020.

Computer Engineering, Israel Institute of Technology, Technion
2. Head Teaching Assistant, Circuit Theory (700+ students), 2015-2018.
3. Head Teaching Assistant, Electronic Switching Circuits (300+ students), 2016-2018.

4. Supervisor of B.Sc. projects, VLSI Lab and Parallel Systems Lab, 2016-2018.
5. Teaching Assistant, MATLAB, 2015-2018.

Department Activity

University of California, Berkeley

1. Graduate PhD Admissions Committee.
2. DARE (Diversifying Access to Research in Engineering) Admissions Committee.
3. Undergraduate project committee.

Advised Students

University of California, Berkeley

1. Chloe Liu (First employment: grad student at Stanford University).
2. Fang Shuo Deng (First employment: software engineer at Abnormal Security).
3. Ian Galbraith (First employment: software engineer at Twilio).

Israel Institute of Technology, Technion

4. Stav Belogolovsky (First employment: DFT engineer at Arbe).
5. Amnon Wahle (First employment: algorithm research at BeyondMinds).

Stav and Amnon won the System Architecture Labs Cluster Prize for an outstanding undergraduate project.

Conference Publications

1. Hasan Genc, Seah Kim, Alon Amid, **Ameer Haj-Ali**, Vighnesh Iyer, Pranav Prakash, Jerry Zhao, Daniel Grubb, Harrison Liew, Howard Mao, Albert Ou, Colin Schmidt, Samuel Steffl, John Wright, Ion Stoica, Jonathan Ragan-Kelley, Krste Asanovic, Borivoje Nikolic, Yakun Sophia Shao, "Gemmini: Enabling Systematic Deep-Learning Architecture Evaluation via Full-Stack Integration," *58th ACM/ESDA/IEEE Design Automation Conference (**DAC 2021**)*, December 2021. **Nominated for Best Paper Award.**

2. Lianmin Zheng, Ruochen Liu, Junru Shao, Tianqi Chen, Joseph E. Gonzalez, Ion Stoica, **Ameer Haj Ali**, "TenSet: A Large-scale Program Performance Dataset for Learned Tensor Compilers," *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track* (**NeurIPS 2021**), December 2021.

3. Lianmin Zheng, Chengfan Jia, Minmin Sun, Zhao Wu, Cody Hao Yu, **Ameer Haj-Ali**, Yida Wang, Jun Yang, Danyang Zhuo, Koushik Sen, Joseph Gonzalez, Ion Stoica, "Ansor: Generating High-Performance Tensor Programs for Deep Learning," *The 14th USENIX Symposium on Operating Systems Design and Implementation (**OSDI 2020**)*, November 2020.

4. Bin Li, Yipeng Wang, Ren Wang, Charlie Tai, Ravi Iyer, Zhu Zhou, Andrew Herdrich, Tong Zhang, **Ameer Haj-Ali**, Ion Stoica, Krste Asanović, "RLDRM: Closed Loop Dynamic Cache Allocation with Deep Reinforcement Learning for Network Function *Virtualization," IEEE Conference on Network Softwarization (**NetSoft 2020**)*, June 2020. **Received Best Paper Award.**

5. **Ameer Haj-Ali**, Qijing Huang, William Moses, John Xiang, John Wawrzynek, Krste Asanović, Ion Stoica "AutoPhase: Juggling HLS Phase Orderings in Random Forests with Deep Reinforcement Learning," *Third Conference on Machine Learning and Systems (**MLSys 2020**)*, March 2020.

6. Keertana Settaluri, **Ameer Haj-Ali**, Qijing Huang, Suhong Moon, Kourosh Hakhamaneshi, Ion Stoica, Krste Asanović, Borivoje Nikolic, "AutoCkt: Deep Reinforcement Learning of Analog Circuit Designs," *Design, Automation & Test in Europe Conference & Exhibition (**DATE 2020**)*, March 2020.

7. **Ameer Haj-Ali**, Nesreen Ahmed, Ted Willke, Sophia Shao, Krste Asanović, Ion Stoica, "NeuroVectorizer: End-to-End Vectorization with Deep Reinforcement Learning," *International Symposium on Code Generation and Optimization 2020 (**CGO 2020**)*, February 2020.

8. **Ameer Haj-Ali**, Nesreen Ahmed, Ted Willke, Sophia Shao, Krste Asanović, Ion Stoica, "End-to-End Vectorization with Deep Reinforcement Learning", *Compiler, Architecture, and Tools Conference 2019 (**CATC 2019**)*, December 2019.

9. **Ameer Haj-Ali**, Qijing Huang, William Moses, John Xiang, Ion Stoica, Krste Asanović, John Wawrzynek, "AutoPhase: Compiler Phase-Ordering for High-Level Synthesis with Deep Reinforcement Learning," *The 27th IEEE International Symposium On Field-Programmable Custom Computing Machines (**FCCM 2019**)*, April 2019.

10. **Ameer Haj-Ali**, Rotem Ben-Hur, Nimrod Wald, and Shahar Kvatinsky, "Efficient Algorithms for In-memory Fixed Point Multiplication Using MAGIC," *IEEE International Symposium on Circuits and Systems (**ISCAS 2018**)*, May 2018.

11. Nishil Talati, **Ameer Haj-Ali**, Rotem Ben-Hur, Nimrod Wald, Ronny Ronen, Pierre-Emmanuel Gaillardon, and Shahar Kvatinsky, "Practical Challenges in Delivering the Promises of Real Processing-in-Memory Machines," *Design, Automation & Test in Europe Conference & Exhibition (**DATE 2018**)*, March 2018.

12. John Reuben, Rotem Ben-Hur, Nimrod Wald, Nishil Talati, **Ameer Haj-Ali**, Pierre-Emmanuel Gaillardon, and Shahar Kvatinsky, "Memristive Logic: A Framework for Evaluation and Comparison," 27^{th} *International Symposium on Power and Timing Modeling, Optimization and Simulation (**PATMOS 2017**)*, September 2017.

Journal Publications

1. Rotem Ben-hur, Ronny Ronen, **Ameer Haj-Ali**, Debjyoti Bhattacharjee, Adi Eliahu, Natan Peled, and Shahar Kvatinsky, "SIMPLER MAGIC: Synthesis and Mapping of In-Memory Logic Executed in a Single Row to Improve Throughput," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (**TCAD**)*, July 2019.

2. Tzofnat Greenberg-Toledo, Roee Mazor, **Ameer Haj-Ali**, and Shahar Kvatinsky, "Supporting the Momentum Training Algorithm Using a Memristor-Based Synapse," *IEEE Transactions on Circuits and Systems I: Regular Papers* (**TCAS-I**), January 2019.

3. **Ameer Haj-Ali**, Rotem Ben-Hur, Nimrod Wald, Ronny Ronen, and Shahar Kvatinsky, "Not in Name Alone: a Memristive Memory Processing Unit for Real In-Memory Processing," ***IEEE Micro***, September 2018.

4. **Ameer Haj-Ali**, Rotem Ben-Hur, Nimrod Wald, Ronny Ronen, and Shahar Kvatinsky, "IMAGING: In-Memory AlGorithms for Image processiNG," *IEEE Transactions on Circuits and Systems I: Regular Papers* **(TCAS-I)**, June 2018.

Book Chapters

1. **Ameer Haj-Ali**, Ronny Ronen, Rotem Ben-Hur, Nimrod Wald, and Shahar Kvatinsky, "Memristor-Based Processing-in-Memory and Its Application On Image Processing," *Elsevier*.

2. Nishil Talati, Rotem Ben-Hur, Nimrod Wald, **Ameer Haj-Ali**, John Reuben, and Shahar Kvatinsky, "mMPU - a Real Processing-in-Memory Architecture to Combat the von Neumann Bottleneck," *Springer*, 2020.

3. John Reuben, Rotem Ben-Hur, Nimrod Wald, Nishil Talati, **Ameer Haj-Ali**, Pierre- Emmanuel Gaillardon, and Shahar Kvatinsky, "A Taxonomy and Evaluation Framework for Memristive Logic," *Springer*, 2017.

Preprints

1. **Ameer Haj-Ali**, William Moses, Qijing Huang, Hasan Genc, John Wawrzynek, Krste Asanović, Ion Stoica, "ProTuner: Tuning Programs with Monte Carlo Tree Search," 2020.

2. **Ameer Haj-Ali**, Nesreen Ahmed, Ted Willke, Joseph Gonzalez, Krste Asanović, Ion Stoica, "A View on Deep Reinforcement Learning in System Optimization," 2019.

Blog Posts

1. Yifei Feng, Sriram Sankar, Siddharth Venkatesh, and **Ameer Haj-Ali**, "Cloud Infrastructure for LLM and Generative AI Applications," 2023.

2. **Ameer Haj-Ali** and Robin Singh "Anyscale Endpoints Preview: Fast, Cost-Efficient, and Scalable LLM APIs," 2023.

3. **Ameer Haj-Ali** and Javier Redondo "Autoscaling clusters with Ray," 2021.

4. **Ameer Haj-Ali**, "Easy Distributed Scikit-Learn with Ray," *Medium*, 2020.

5. **Ameer Haj-Ali**, "Scale ML on Your Local Clusters with Ray ," *Medium*, 2020.

Invited Talks

1. Nvidia GTC, March, 2024.
2. Fine-Tuning and Evaluating LLM Retrieval Models with Anyscale and Arize meetup, Dec, 2023.
3. Ray Summit, September, 2023.
4. Generative AI Summit Silicon Valley, September 2023
5. Intel Labs, June, 2020.
6. Huawei Technologies, May, 2020.
7. Google Brain, May, 2020.
8. VMware research, February, 2020.
9. Technion, January, 2020.
10. The Hebrew University of Jerusalem, December, 2019.
11. Intel Corporation Israel, December, 2019.
12. Intel Labs, August, 2019.

Workshops & Events Co-organized

1. Stephen and Sharon Seiden Frontiers in Engineering and Science Workshop: Beyond CMOS, From Devices to Systems, Haifa, Israel, 2017.
2. Directed the first Ray meetup 2019 in Israel, Tel-Aviv.

Posters

1. **Ameer Haj-Ali**, Nesreen Ahmed, Ted Willke, Sophia Shao, Krste Asanović, Ion Stoica, "Learning to Vectorize Using Deep Reinforcement Learning", *Workshop on ML for Systems at NeurIPS 2019*, December 2019.

2. **Ameer Haj-Ali**, Qijing Huang, John Wawrzynek, Ion Stoica, and Krste Asanović, "Using Deep Reinforcement Learning for Compiler Optimization Selection and Ordering," *Intel Summit (Santa Clara)*, October 2018.

3. **Ameer Haj-Ali** and Shahar Kvatinsky, "Beyond von-Neumann Computing," *Technion Research Day*, November 2017.

4. **Ameer Haj-Ali**, Rotem Ben-Hur, Nimrod Wald, and Shahar Kvatinsky, "In-memory Image Processing Algorithms," *Beyond CMOS: From Devices to Systems*, June 2017.

5. **Ameer Haj-Ali**, Nimrod Wald, and Shahar Kvatinsky, "Configurable Simulator for Multithreaded Processors," *Intel Collaborative Research Institute*, May 2017.

Conference & Journal Referee

NeurIPS 2019, HPCA 2018, DATE 2018, VLSI-SoC 2018, ISCAS 2017, ISCAS 2016, CNNA 2016. IEEE Transactions on Circuits and Systems I (TCAS-I), IEEE Transactions on Circuits and Systems II (TCAS-II), IEEE Transactions on Very Large Scale Systems (TVLSI), Microelectronics Journal.

Community Service

1. Area Chair: NeurIPS 2022, 2023 Track Datasets and Benchmarks.
2. On the judge panel of Tsofen Ultra Hack 2.0 Hackathon 2020.
3. Promoting engineering within the underprivileged Arab community in Israel: giving lectures and appearing on multiple radio programs to raise awareness and inspire motivation for higher education in the Arab community, 2017-present.
4. Distributing food on holidays, 2015-present.
5. "Project for Equal Opportunities": accompany new minority students at the Technion to help them integrate into the new curriculum, including workshops, personal assistance, exam preparation and motivation, 2014-2018.
6. Represent my school at a conference in Italy (Udine) where students from many races and religions participated for endowments and promoting diversity, 2012.
7. First aid in "Noar Maccabi" youth program, 2011-2012.

Languages and Skills

Classical/Colloquial Arabic, Hebrew, English.
Python, TensorFlow, PyTorch, vLLM, Ray, Android, JavaScript, CUDA, C, C++, C#, UNIX, XML, MATLAB, HTML, Regular Expressions, Assembly Language, SQL, Bluespec, Chisel, Verilog, Verilog-A, VHDL, Cadence Virtuoso, Synopsys Design Vision, Cadence Encounter, PSpice, NVSim, NVMain, Nvidia-smi, LaTeX.