# AARON HARLAP

aaron.harlap@gmail.com

## OVERVIEW

**Determined AI, Boston, MA.**

Senior Software Engineer — May 2020 - Current
Software Engineer — June 2019 - May 2020

**Carnegie Mellon University, Pittsburgh, PA.**

Ph.D., Electrical and Computer Engineering — May 2019

- Advisors: Greg Ganger and Phil Gibbons
- Research Topic: Large Scale Machine Learning in Shared Computing Environments

Master of Science, Electrical and Computer Engineering — May 2016

**Northeastern University, Boston, MA.**

Bachelor of Science, Electrical and Computer Engineering — May 2014

## EXPERIENCES

Experienced with working at the intersection of systems and machine learning with a focus on distributed systems. Designed and built production grade machine learning platforms for a variety of uses cases including: autonomous self driving, drug research, and fraud detection. Experienced with leading projects for small and medium sized teams, and leading architecture groups.

**Determined AI** — June 2019 - Current

- Worked on the Determined training platform: https://github.com/determined-ai/determined.
- Led high impact projects including: distributed training, data layer, and Kubernetes.
- Improved product capability from no support for distributed training to 2x faster than competing solutions.
- Led the design and implementation of re-architecting the platfrom to integrate with Kubernetes.
- Led infrastructure architecture group that drove sprint planning and long term technical vision.
- Open sourced YogaDL to improve data input for Tensorflow: https://github.com/determined-ai/yogadl.

| | |
|---|---|
| **Programming Languages:** | Python, Go, C++ |
| **Software Tools:** | Git, Docker, Circle CI, Helm |
| **ML Frameworks:** | PyTorch, TensorFlow |
| **Platforms:** | Kubernetes, AWS, GCE |

## PHD RESEARCH

**Thesis**: Improving efficiency, run-time and cost of machine learning applications in cloud environments.

- **Committee**: Greg Ganger, Phil Gibbons, Ameet Talwalkar, and Amar Phanishayee

**PipeDream: Pipeline Parallelism for DNN Training**

*Published at SOSP'19*

- Designed PipeDream, an efficient data-parallel+model-parallel system for distributed deep learning.
- Achieved 5x faster DNN training as compared to prior techniques.
- Achieved near linear scaling on modern hardware (v100 GPUs) where prior techniques struggled.
- Integrated with PyTorch, a popular deep learning framework.
- Open sourced at: https://github.com/msr-fiddle/pipedream.

**Tributary: spot-dancing for elastic services with latency SLOs**

*Published at Usenix ATC'18*

- Designed Tributary, a system for running services with latency SLOs on pre-emptible resources.
- Built and deployed a machine learning model for predicting pre-emption of Amazon EC2 Spot instances.
- Developed a cost-model for acquiring resources in order to meet user specified SLO requirements.

- Experimented with real-world web-service traces, and observed cost savings up to 85% for achieving same SLOs compared to using non-preemptible resources.

**Proteus: agile ML elasticity through tiered reliability in dynamic resource markets.**
*Published at EuroSys'17*

- Designed Proteus, a agile elastic machine learning system that efficiently runs on pre-emptible instances.
- Proposed new parameter-server architecture to efficiently handle bulk resource pre-empetion.
- Implemented a novel resource manager for Amazon EC2 that decreased cost for ML applications by 85%.
- Experimented with real ML tasks, running on pre-emptible Amazon EC2 instances.

**Addressing the straggler problem for iterative convergent parallel ML**
*Published at SoCC'16*

- Observed adverse straggler effects on ML training systems running on Amazon EC2 and Microsoft Azure.
- Designed a parameter server system that supports temporary work-reassignment and relaxed worker synchronization.
- Experimented with many real ML applications, running on Amazon EC2 and Microsoft Azure, observing improvements up to 3x over prior approaches.

## INTERNSHIPS

**Microsoft Research** May 2017 to Aug 2017
*Research Intern*

- Developed research ideas and system implementation for a novel machine learning training system.
- Incorporated into PhD thesis (*PipeDream* project).
- Published at *SysML' 18* and *SOSP' 19*.

## PUBLICATIONS

1 **Aaron Harlap**, Deepak Narayanan, Amar Phanishayee, Vivek Seshadri, Gregory R. Ganger, Phillip B. Gibbons. PipeDream: Pipeline Parallelism for DNN Training. In *ACM Symposium on Operating Systems Principles (SOSP'19)*, 2019.

2 **Aaron Harlap**, Andrew Chung, Alexey Tumanov, Gregory R. Ganger, Phillip B. Gibbons. Tributary: spot-dancing for elastic services with latency SLOs. In *USENIX Annual Technical Conference (Usenix ATC' 18)*, 2018.

3 **Aaron Harlap**, Alexey Tumanov, Andrew Chung, Gregory R. Ganger, Phillip B. Gibbons. Proteus: agile ML elasticity through tiered reliability in dynamic resource markets. In *ACM European Conference on Computer Systems (EuroSys'17)*, 2017.

4 Kevin Hsieh, **Aaron Harlap**, Nandita Vijaykumar, Dimitris Konomis, Gregory R. Ganger, Phillip B. Gibbons, Onur Mutlu. Gaia: Geo-Distributed Machine Learning Approaching LAN Speeds. In *USENIX Symposium on Networked Systems Design and Implementation (NSDI' 17)*, 2017.

5 **Aaron Harlap**, Henggang Cui, Wei Dai, Jinliang Wei, Gregory R. Ganger, Phillip B. Gibbons, Garth A. Gibson, and Eric P. Xing. Addressing the Straggler Problem for Iterative Convergent Parallel ML. In *ACM Symposium on Cloud Computing (SoCC'16)*, 2016.

## OTHER EXPERIENCES

| | |
|---|---|
| Invited speaker at Machine Learning Meetups in NY and Boston. | 2020 |
| Invited panelist at the Vector Institute ML Workshop. | 2019 |
| Captsone Advisor. | 2018 |
| ECE Department organizer for basketball and softball teams. | 2015-2018 |
| Teaching Assistant 15746/18746 Storage Systems. | 2016-2017 |