AARON HARLAP

aaron.harlap@gmail.com

EDUCATION

Carnegie Mellon University, Pittsburgh, PA.

Ph.D., Electrical and Computer Engineering (GPA: 3.89/4)   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 (GPA: 3.87/4)   May 2014

SKILLS

Programming languages	C++, Python
Software Systems	Git, Linux, Shell, LaTeX
Big data Systems	PyTorch, TensorFlow, Caffe
Platforms	AWS, GCE

PHD THESIS RESEARCH

Thesis: Improving efficiency, run-time and cost of machine learning applications in cloud computing environments
- **Committee**: Greg Ganger, Phil Gibbons, Ameet Talwalkar, and Amar Phanishayee
- **Thesis Subprojects**: As follows:

PipeDream: Pipeline Parallelism for DNN Training
Published at SysML'18
- Designed PipeDream, an efficient data-parallel+model-parallel system for distributed deep learning.
- Extended Caffe, a popular deep learning system, to run on distributed GPU machines, by using PipeDream.
- Achieved 5x faster DNN training as compared to prior techniques.
- Achieved near linear scaling on modern hardware (v100 GPUs) where prior techniques struggled.

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*.

Spectral Sciences Incorporated   January 2013 to June 2013
Co-Op Software Engineer
- Developed graphical interface for new applications delivered to customer.
- Developed and implemented tests for existing code base.
- Completed Modtran code upgrade process, removing Modtran 77 dependencies.

Motorola Mobility   January 2012 to June 2012
Co-Op Software Engineer
- Developed early patching system for the BSR sustaining team.
- Implemented IPV6 neighbor discovery customer interface.

Charles River Development   January 2011 to August 2011
SQA Co-Op Engineer
- Developed automated resource status check system for the automaton and infrastructure group.
- Rebuilt and centralized regression testing suite, reducing testing time by over 80%.

PUBLICATIONS

1. **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.

2. **Aaron Harlap**, Deepak Narayanan, Amar Phanishayee, Vivek Seshadri, Gregory R. Ganger, Phillip B. Gibbons. PipeDream: Pipeline Parallelism for DNN Training. In *SysML (SysML'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.

COURSES

18847 Machine Learning Infrastructure	15719 Advanced Cloud Computing
10701 Machine Learning	15721 Advanced Database Systems
18749 Building Reliable Distributed Systems	18601 Entrepreneurship Innovation Technology
18746 Storage Systems	15712 Advanced Topics in Operating Systems

LEADERSHIP EXPERIENCE

Captsone Advisor.	2018
ECE Department organizer for basketball and softball teams.	2015-2018
Teaching Assistance: 15746/18746 Storage Systems.	2016-2017