Sung-Han Lin

sunghlin@gmail.com | +1-213-810-2139 | https://sunghlin.github.io/about

Education

Ph.D. in Computer Science - *University of Southern California*, Los Angeles, California, U.S.A. 2010 – 2017

- **Dissertation**: Distributed Resource Management for QoS-Aware Service Provision
- Teaching Assistant, **CSci 402: Operating Systems**, Fall 2012 - Summer 2017

B.S. and M.S. in Computer Science - *National Taiwan University*, Taipei, Taiwan 2002 – 2008

Selected Work Experience

Performance Analysis Engineer & Data Scientist - *NetApp Inc.* October 2017 - Present

- **Improving GPU Data Pipeline**: [*Python*, *TensorFlow*, *Horovod*, *Container*, *MPI*, *Kubernetes*, *S3*]
  - Designed and Deployed the AI/ML/DL reference architecture combining NVIDIA DGX systems and NetApp storage systems for distributed TensorFlow training and inference over the RoCE connections.
  - Identified the bottleneck of the end-to-end data pipelines; improved the GPU utilization, data feeding rate, and training speed via better configuring the CPU workload and multithreading, GPU memory usage, and network interconnections, mainly focusing on image recognition training (ResNet, VGG, Inception).
  - Adopted NVIDIA NGC and libraries to our reference architecture; implemented and optimized ML/DL models with various frameworks and published Technical Report for different verticals - autonomous vehicle (Mask-R-CNN), financial (Semi-supervised GAN, Auto-Encoder), and conversational AI (BERT).
- **Data Transmission Failover**: Built the automation and investigated the executing path of failover, where the secondary clusters takes over the write operations from the primary cluster; helped reduce the latency via improving the logic of triggering failover and the tie breaker. [*C/C++*, *Python*]
- **Synchronous Replication**: Built the automation and investigated the latency of replicating write operations to the remote cluster synchronously; helped reduce the overhead of transmission via identifying the bottleneck and improved the concurrency of parallel operations. [*C/C++*, *Tcl*, *Python*]
- **Advanced Block Fetching**: Investigated the block prefetching strategies to reduce the read overhead and latency caused by the checksum mechanism of regular disks; improved the read throughput by 20%. [*C/C++*]

Intern - *Teradata* June 2014 - August 2014

- **Improving SQL-MapReduce Engine**: Re-designed and implemented the data buffering and transmission mechanism to improve the query throughput by 15% via reducing the I/O latency and function overhead between user defined functions and databases. [*C/C++*, *Hadoop*, *SQL*]

Selected Publications

- **A Model-based Approach to Streamlining Distributed Training for Asynchronous SGD**, by *Sung-Han Lin*, Marco Paolieri, Cheng-Fu Chou, Leana Golubchik, IEEE MASCOTS, 2018
- **Performance Driven Resource Sharing Markets for the Small Cloud**, by *Sung-Han Lin*, Ranjan Pal, Marco Paolieri, Leana Golubchik, IEEE ICDCS, 2017
- **Sustaining Ad-Driven P2P Streaming Ecosystems A Market-Based Approach**, by *Sung-Han Lin*, Ranjan Pal, Bo-Chun Wang, Leana Golubchik, IEEE/ACM IWQoS, 2015

Selected Research Projects

- **Throughput Estimation for Large-scale Deep Learning**: Analyzed the traffic pattern of Asynchronous SGD training in the parameter-server architecture, and built a queueing network model to estimate the training speed (images processed per second) of distributed TensorFlow GPU systems; leveraged the estimation model to address the problem of scheduling heterogeneous distributed DNN training jobs in a computing cluster.

Talk

- **Performance Considerations for AI and ML Deployments**, NetApp INSIGHT Las Vegas, October 2019