Sung-Han Lin

sunghanl@fb.com | +1-213-810-2139 | sunghlin.github.io

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 (CSIE) - *National Taiwan University*, Taipei, Taiwan 2002 – 2008

- Master Thesis: **On the Design of Vehicular P2P Scheme over Ad Hoc Network and the Internet**

Selected Work Experience

Performance and Capacity Engineer - *Facebook* January 2021 - Present

- **Improving Inference Performance:** [*Python*, *Caffe2*, *PyTorch*]
  - Migrating the inferencing model from CPU to the accelerators

Performance Analysis Engineer & Data Scientist - *NetApp Inc.* October 2017 - January 2021

- **ONTAP AI: Improving GPU Data Pipeline:** [*Python*, *TensorFlow*, *Horovod*, *Container*, *MPI*, *Kubernetes*, *S3*]
  - Designed and Deployed the 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, the GPU memory usage, and network interconnections, mainly on image recognition training and inference (*ResNet*, *VGG*, *Inception*).
  - Adopted NVIDIA NGC and libraries to build our reference architecture; implemented and optimized ML/DL training models, and published joined Technical Report with NVIDIA on various verticals - autonomous vehicle (*Mask-R-CNN*), financial (*Semi-supervised GAN*, *Auto-Encoder*), and conversational AI (*BERT*).
- **Data Operation Failover**: Built the automation and investigated the executing path of failover, where the secondary cluster takes over the write operations from the primary cluster; helped reduce the I/O resumption latency via improving the logic of triggering failover and reducing the function overhead. [*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 improving 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 new checksum mechanism for disks; improved the read throughput by 20%. [*C/C++*]

Intern - *Teradata* June 2014 - August 2014

- **Improving SQL-MapReduce Execution Engine:** Redesigned and implemented the data buffering and transmission mechanisms 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

- **On the Economic Sustainability of Cloud Sharing Systems Are Dynamic Single Resource Sharing Markets Stable?**, Ranjan Pal, Aditya Ahuja, *Sung-Han Lin*, Abhishek Kumar, Leana Golubchik, Nachikethas A Jagadeesan, ACM SIGMETRICS Performance Evaluation Review, 2019
- **Are Federated Cloud Sharing Systems Sustainable?: On Dynamic Sharing Markets and Their Stability**, Ranjan Pal, Aditya Ahuja, *Sung-Han Lin*, Nachikethas Jagadeesan, Abhishek Kumar, Leana Golubchik, IEEE Transactions on Sustainable Computing, 2019
- **A Model-based Approach to Streamlining Distributed Training for Asynchronous SGD**, *Sung-Han Lin*, Marco Paolieri, Cheng-Fu Chou, Leana Golubchik, IEEE Symposium on Modelling, Analysis, and Simulation of Computer and Telecommunication Systems (MASCOTS), 2018

- **Performance Driven Resource Sharing Markets for the Small Cloud**, *Sung-Han Lin*, Ranjan Pal, Marco Paolieri, Leana Golubchik, IEEE International Conference on Distributed Computing Systems (ICDCS), 2017
- **On a Market-Driven Hybrid P2P Video Streaming Approach**, *Sung-Han Lin*, Ranjan Pal, Bo-Chun Wang, Leana Golubchik, IEEE Transactions on Multimedia, 2017, Issue Date: May.2017, Volume: 19, Issue: 5, pages: 1-15
- **Sustaining Ad-Driven P2P Streaming Ecosystems A Market-Based Approach**, *Sung-Han Lin*, Ranjan Pal, Bo-Chun Wang, Leana Golubchik, IEEE/ACM International Symposium on Quality of Service (IWQoS), 2015

Patents

- **Network resource allocation system and method of the same**, Cheng-Fu Chou, Ching-Ju Lin, *Sung-Han Lin*, US Patent 8,116,324

Talk Presentations

- **Data Pipeline and Performance Considerations for NetApp AI**, NetApp INSIGHT Digital Event, 2020
- **AI Infrastructure for Real-World Use Cases from NetApp and NVIDIA (Presented by NetApp)**, GPU Technology Conference (GTC), 2020
- **Performance Considerations for AI and ML Deployments**, NetApp INSIGHT Las Vegas, 2019

Skills

- Proficiency in software development in C/C++, Perl, Python, Shell Scripts, HTML/CSS, JavaScript, Tcl, JAVA
- Familiarity with TensorFlow, Horovod, Container, MPI, Kubernetes, SQL, S3, Caffe2, PyTorch, Hadoop, Sockets, Matlab, PHP, UNIX, GNU/Linux, Solaris