Chang-Hong Hsu

Senior Software Engineer & Technical Lead

bnsblue@gmail.com | +1-(734)2392423 | Greater Seattle area, WA, USA

Profile

- A senior software engineer and a technical lead at Lyft, Inc. with profound experience spanning backend and ML model productionization:
  - Supervised and engineered the production systems hosting Lyft's critical ML models that predict rider conversion signals and driver supply forecasts, and the cross-functional integration with pricing products.
  - Developed critical features in Flyte to improve robustness and add functionalities for massive-scale data processing and ML model training in an multi-tenant-multi-cluster environment.
  - Engineered a deep-learning-based in-cabin Perception stack and developed applications facilitating a friendlier and safer riding experience in autonomous vehicles.
- A PhD in Computer Engineering with hands-on & research experience on applied machine-learning / deep-learning, and cross-layer acceleration and optimization for datacenter and ML system efficiency.
- Passionate about every opportunity sitting at the intersection of systems and machine learning. Eager to design and implement large-scale and efficient machine learning systems, and leverage knowledge of computer architecture and system architecture to improve latency, throughput, and efficiency for such systems.

Skills

Programming Languages: C++, Python, Golang, Protobuf, Java, Verilog
Container & orchestration: Docker, Kubernetes, Flyte (open-source contributor), Airflow
Tools and frameworks: Apache Beam (Streaming), TensorFlow, Keras, Caffe, OpenCV, scikit-learn
Cloud: AWS (EC2, SageMaker, Cloudwatch, ECS, ECR, Array, etc.)

Work Experience

Senior Software Engineer and Technical Lead, Lyft Inc. (Jul. 2018 – now)

Rider conversion and real-time supply forecasting model

- Tech-leading and supervising the engineering and **productionization of Lyft's business-critical long- and short-term rider conversion prediction models**; led the cross-functional collaboration and integration with Lyft's core pricing products to generate $10'sM+ expected financial impact
- Led the implementation of a DNN-based **real-time supply forecasting model and pipeline**, which achieved a 2-hour avg Neighborhood L1 MASE ~= 0.8 and an avg. bias < 5%
- Leading the engineering of **the rider conversion models & the streaming-based online bias correction module with Apache Beam; leading the engineering of the life cycle such as the serving, training infrastructure, and backtesting framework** for the model

Flyte: an open-source, k8s-native, large scale orchestration engine for data processing and machine learning (https://github.com/lyft/flyte)

- **Led the integration of AWS SageMaker on Flyte**, allowing users to access and leverage SageMaker's distributed Training and Hyper Parameter Tuning capabilities directly from a Flyte workflow
- Significantly improved Flyte robustness by (1) **facilitating fairness and the progress guarantees for resource allocation** (e.g., #each tenant's outstanding Trino requests) for a multi-cluster, multi-tenant compute environment, (2) **optimize memory consumption** for large dataframe manipulation, and (3) implement **intelligent back-off mechanisms for pod creation**
- Impact: Reduced resource-contention related pages by >90%. Reduced memory footprint for 40% – 75% when processing 10's-GB-scale dataframes. Reduced 75% Pod Creation calls to Kubernetes' API server

In-cabin Perception for Autonomous Vehicle

- Designed and engineered an end-to-end **in-cabin perception solution for autonomous vehicles for ride-share** to facilitate applications such as cabin tidiness checking, left-behind objects detection, etc. This includes the inference stack and the apparatus, the software pipeline, and the environment for low-cost, highly flexible and portable in-cabin data collection.
- Designed and implemented a luminance-based frame sampling algorithm to automatically extract frames with distinct lighting conditions from a pool of frames which balanced the highly-skewed training set.
- Adapted and trained ResNet-50, YOLO-9000, and AnoGAN to check cabin tidiness and detect left-behind objects. Achieved up to 0.92 AUC ROC and up to 91% accuracy for the tidiness checker

Selected Internship Experience

Research Intern, Facebook Inc.

- Designed a datacenter power-capping runtime and validate the design decisions
- Developed a framework that helps derive service placement for highly efficient power budget utilization

Education

Ph.D., Computer Science and Engineering, University of Michigan, Ann Arbor 2012 – 2018

- Thesis title: **Towards Power- and Energy-efficient Datacenters**
- Co-advisers: Prof. Jason Mars and Prof. Lingjia Tang

M.S., Electrical Engineering, National Taiwan University 2008 – 2011

B.S., Electrical Engineering, National Taiwan University 2004 -- 2008

Selected Academia Projects

Architectural and System Implication of Accelerated Video Analytic for Dash-cam Videos

- Integrated deep learning pipelines to analyze video and sensor data and answer complex user queries
- Proposed a feature-reusing-based optimization to accelerate Convolutional Neural Network-based (CNN-based) video analytic algorithms; leveraged AVX instructions to accelerate GEMM computation on coarse-grain frame blocks by up to 9x

Architecture and Interface of Future Intelligent Vehicles

- Investigated the constraints and acceleration options for autonomous vehicles -- explored different algorithms and acceleration platforms in the visual pipeline of such a system; achieved up to 163x reduction in end-to-end tail latency
- Constructed a voice-triggered on-vehicle digital assistant taking human commands in natural language to check vehicle status and manipulate states of certain safety-related vehicle features -- defined and configured OBD-II interface with OpenXC to probe and write signals from/to the vehicle CAN bus

Addressing DNN execution bottleneck on GPUs

- Addressed DNN execution bottleneck on GPUs by (1) dropping the non-contributing synapses in the neural network, and (2) reducing the data movement requirements within DNN computations
- Improved performance by over 2.1× on commodity GPU hardware and over 2.6× when leveraging a small additional custom hardware unit

Reducing Power Budget Fragmentation Problem in Large-Scale Datacenters

- Identified root cause of suboptimal power budget utilization in large-scale datacenters
- Leveraged temporal heterogeneity of the power consumption patterns among different services and designed a clustering-based service placement framework to optimize power utilization
- Increased the throughput under a fixed power constraint by up to 8% in production data centers significantly without changing the power-delivery infrastructure

Pinpointing and Reining in Tail Queries with Quick Voltage Boosting

- Developed a framework that improves tail latency of important cloud applications (Memcached and Web Search) by more than 4× with high energy efficiency
- Identified query-level indicator to predict and pinpoint tail queries, and design low-overhead DVFS policy to utilize quick voltage switching circuits to boost system performance for tail queries

Selected Publications

Google Scholar Citations (#citations = 970 as of 4/24/23): https://scholar.google.com/citations?user=gZzj1-8AAAAJ&hl=en

Designing and Optimizing Machine Learning and Deep Learning Systems

[UIST'18] Adasa: A Conversational In-Vehicle Digital Assistant for Advanced Driver Assistance Features
S.-C. Lin, C.-H. Hsu, W. Talamonti, Y. Zhang, S. Oney, J. Mars, L. Tang. (UIST'18 Honorable Mention Award)

[ASPLOS'18] The Architectural Implications of Autonomous Driving: Constraints and Acceleration
S.-C. Lin, Y. Zhang, C.-H. Hsu, M. Skach, M. Haque, L. Tang, J. Mars

[MICRO'17] DeftNN: Addressing Bottlenecks for DNN Execution on GPUs via Synapse Vector Elimination and Near-compute Data Fission
P. Hill, A. Jain, M. Hill, B. Zamirai, M. Laurenzano, C.-H. Hsu, S. Mahlke, L. Tang, J. Mars

Cross-layer Optimization for Modern Datacenters

[ASPLOS'18] SmoothOperator: Reducing Power Fragmentation and Improving Power Utilization in Large-scale Datacenters
C.-H. Hsu, Q. Deng, J. Mars, L. Tang

[IEEE IC] Thermal Time Shifting: Decreasing Datacenter Cooling Costs with Phase Change Materials
M. Skach, M. Aurora, C.-H. Hsu, Q. Li, D. Tullsen, L. Tang, J. Mars

[ACM TOCS] Achieving Short Tail Latency with High Energy Efficiency for Warehouse-scale Computers with Adrenaline
C.-H. Hsu, Y. Zhang, M. A. Laurenzano, D. Meisner, T. Wenisch, R. G. Dreslinski, J. Mars, L. Tang

[ISCA'16] Dynamo: Facebook's Data Center-Wide Power Management System
Q. Wu, Q. Deng, L. Ganesh, C.-H. Hsu, Y. Jin, S. Kumar, B. Li, J. Meza, Y. J. Song

[ISCA'15] Thermal Time Shifting: Leveraging Phase Change Materials to Reduce Cooling Costs in Warehouse-Scale Computers
M. Skach, M. Arora, C.-H. Hsu, D. Tullsen, J. Mars, L. Tang

[HPCA'15] Adrenaline: Pinpointing and Reining in Tail Queries with Quick Voltage Boosting
C.-H. Hsu, Y. Zhang, M. A. Laurenzano, D. Meisner, T. Wenisch, J. Mars, L. Tang, R. G. Dreslinski