# Jasper **Shan**

ML Infrastructure / Training Systems Engineer

(+1) 669-282-0848 | jaspershan@gmail.com | dollarplus

## Skills

**Programming Languages** Python, C++, CUDA, Java

**Libraries / Frameworks** PyTorch, NCCL, Triton, TorchRec, FBGEMM, Django

**Systems / Techniques** Distributed Training, Model/Data Parallelism, GPU Memory Optimization, Performance Profiling

## Work Experience

### Meta

*Menlo Park, California*

Senior Software Engineer | Training Optimizations, TorchRec — *July 2023 - Now*

- Led the Training Optimizations team to land training optimization techniques - including embedding pruning, embedding offloading, and model freshness, saving Meta upwards of $250M per year in compute costs.
- Drove embedding pruning techniques that reduce embedding table sizes by up to 50%, significantly lowering serving and training capacity requirements, saving up to 2 MW of power.
- Drove adoption of embedding offloading across Meta's leading recommendation models, enabling frontier-scale training at 100TB by leveraging GPU memory optimization with host memory and SSD tiers, and integrating with 2D sparse parallelism across multi-node GPU clusters.
- Reduced model staleness from ~30 minutes to ~5 minutes for Ads and Recommendation Models by driving model freshness and Early Stage Ranking initiatives.
- All techniques contributed upstream to the open-source TorchRec and FBGEMM libraries on PyTorch.

Software Engineer | Messenger Instagram Direct — *February 2021 - July 2023*

- Drove efforts around the migration of the entire messaging stack from Django (Python) to WWW (Hack)
- Led efficiency initiatives around message rendering which resulted in a total of 1 MW of power savings
- Implemented user tooling around message rendering to improve DevX efficiency. Similar workflows saw 90% reduction in engineering hours required for task completion.

Software Engineer | Privacy Infra — *January 2019 - February 2021*

- Led projects around a company-wide data classification system to categorize and label all stored data at Meta, enforcing safety and compliance guardrails at infrastructure scale.
- Designed and implemented a real-time streaming classification system for sensitive data that cannot be persisted, handling all of Meta's inbound and outbound traffic.
- Trained custom ML models to classify and label free-form user data types, ensuring compliance with privacy requirements across billions of records.

### Facebook

*Menlo Park, California*

SWE Intern | Instagram Privacy & Security — *Jan 2018 - April 2018*

- Developed an Audience Control framework that allows developers to easily specify a custom audience (Followers, Following, Close Friends) for any combination of interactions, and content type on Instagram.

### Amazon

*Seattle, Washington*

SWE Intern | AWS Lambda — *May 2017 - December 2017*

- Implemented new feature for AWS Lambda API which enables Optimistic Concurrency Control around Lambda Function updates, allowing developers to guarantee the consistency of their function update transactions.

## Education

### University of Waterloo

*Waterloo, Ontario*

B.S. in Computer Science — *Sept. 2013 - August 2018*