Peiyuan Qi

+1 (213) 255-6802 | peiyuanqi@gmail.com | Linkedin

PROFESSIONAL EXPERIENCE

Google LLC. (Google) Sunnyvale, CA

Software Engineer *– Google Cloud - ML, Systems, Cloud AI (MSCA).* *Jun. 2021 - Present*

- **Managed** multiple software programs as program Tech Lead of the domain services in the team, for data center modeling.
- **Launched** multiple successful cross-organizational programs without outages, which outcomes are featured in external documents and news: GKE Compact Placement, TPU v5e, A3 GPU, Trusted Partner Cloud program, saving the company avg $XX million per projects, and opening market of $X billions, cooperated closely with principle engineers. Reacted quickly with simultaneous business demands for New Product Introductions and pushed multiple projects to progress at the same time.
- **Led** multiple task forces (six engineers) for design and implementation from datacenter supply chain software in programs.
- Led the **key transformations** of Google's GPU Network deployment strategy to better meet the hardware needs of NCCL, from A3 to more powerful GPU machine racks, with high level design, project requirement doc, engineering reviews.
- Defined the **API** following Google AIP and **SLO** in the perspectives of availability and latency for backend services.
- Sympathized with stakeholders (ML, Cloud customers) in need of the tooling to fulfill the business needs, by actively interviewing the customers, with **product descriptions doc**, business contracts, and operation protocols.
- Drove the solution of data consistency issues cross orgs, and solve operational problems with classic algorithms, i.e. **Hungarian**.
- Acted as the go-to person in team for knowledge about Google infra tools, in term of **ACL, metrics, monitoring, alerting**.
- Crafted the **math solver** (CP-SAT/MIPS) modeling problem based on constraints in for **Power, Network, Space, Cooling**.
- Maintained the business-critical services with service (1M+ lines of business logic code) debugs, **unit tests, integration tests,** release rollouts, Database managements. Eased with the explosion of system complexity with dev guidance and code validators.
- Engineered the **asynchronized** task management system in GUI to improve user experience using internal workflow framework.
- Transformed the existing DOM object rending method to WebGL with **Entity Component System** model using **threejs** as a team, with **Data Oriented Design** principle. Developed frontend features with **Angular and Typescript**, backend features with Python and Java servers. Provided monitoring, metrics reporting and alerting supports with internal tools.
- Explored the possibility of employing **Reinforcement Learning** model to replace the math solver, Embedding models for data center planning entities (orders, forecast data, machine shapes, rack physical data, network demands, etc) representations.
- **Mentored** new team members since 2022, hosted interns, who got Google return offers and joined in 2024. Projects include using internal **LLM API** to build RPC for non-tech PM to use team services, building full-stack web features.

Electronic Arts Inc. (EA) Redwood City, CA

Software Engineer *– EA Data & AI Platform* *Feb 2020 – Jun 2021*

Software Engineering Internship *May 2019 – Aug 2019*

- Developed the container images for legacy Spring Boot socket server and laid out the infra to onboard the app into AWS EKS as pathfinding engineer in the org. And conducted the stress test up to 5k QPS to ensure the robustness and the autoscaling ability.
- Overcame the technical obstacles by communication with opensource contributors of load balancing middleware.
- Built the service monitoring, business reporting, and alerting with the KairosDB, CloudWatch, Elasticsearch, Kibana.
- Migrated EA data gateway service from cloud instances into AWS EKS with autoscaling to save 50% current budget.
- Owned the Kafka and Apache Storm for the real-time metrics processing supporting the billing of the games.
- Worked on the migration of the Apache Storm cluster into the Flint to unify the real-time and batch data pipeline.

Interned at NVIDIA Shanghai, Intel Shanghai *Aug 2017 – Aug 2018*

- Worked on RESTful API design and command line tool delivery in open-source community.
- Applied doc2vec and WMD algorithms to perform the Bug Triage tasks to improve team debugging efficiency.

OTHER PROJECTS

A Fine-Tuned Language Model based on LLaMA2 with personal blog posts with TRL SFT and PEFT LoRA	Dec 2023
A program running YoloV3 on Raspberry Pi with Camera for object detection.	2021
A Radio-Controlled Sailboat with Autonomous Navigation (Frsky, OpenTX, ArduPilot)	Nov 2021

EDUCATION

University of Southern California (USC), Los Angeles, CA

Master of Science in Computer Science *Aug 2018 – Jan 2020*

Shanghai Jiao Tong University (SJTU), Shanghai, China

Bachelor of Science in Electrical and Computer Engineering Graduate with Honors *Sept 2014 – Aug 2018*

SKILLS

Python, Java, gRPC, Angular, Typescript, PyTorch, JAX, Transformers, Streaming Pipeline, Kafka, Storm, Zookeeper, Kubernetes, Docker, Nginx, Spring, Hadoop, Prometheus, C++, Figma. Google Internal Tools (Service Platform Boq/Pod, Stubby, PubSub, Spanner, Flume, Gaia, Ganpati, MDB, UberProxy, Stubby, Spanner, Plx, Sigma, Analog, Automon, SilkRoads).