Full-Time Job

Principal Engineer, Cloud ML Compute Systems - Sunnyvale

About this role

Full-Time Job (Level 8)

Location: Kirkland, Sunnyvale

Job family: Software Engineering

Job subfamily: Software Engineering

Product Area: GCloud GCP

Updated Oct 7 | 279 views

Team members

Newfel Harrat
US_REMOTE
Engineering Director (hiring...

James Chamberlain
Mountain View
Recruiter, Leadership Staffing

Description

Google Cloud is searching for a highly skilled and motivated Principal Engineer to optimize machine learning model performance for our customers and help them achieve maximum model performance for large scale training and inference through tuning and optimization at both software and hardware levels. In this pivotal role, you will collaborate closely with customers, write custom kernels, and develop custom solutions to meet their unique model performance requirements. A deep understanding of deep learning frameworks (like PyTorch or JAX), strong coding skills, excellent communication abilities, and a passion for mentoring junior engineers are essential for success in this role.

Minimum Qualifications:

- Bachelor’s degree in Computer Science, or similar technical field of study or equivalent
- 15 years of professional experience as a software engineer or 13 years with an advanced degree.
- Experience working with GPUs and other hardware accelerators
- Deep expertise with ML Frameworks (PyTorch or JAX or TensorFlow)

Preferred Qualifications:

- 20 years of professional experience.
- Experience in optimizing machine learning models
- Experience writing custom kernels (CUDA, Pallas, etc) is highly desired
- Experience with ML workload performance profiling and analysis
- Strong programming skills in Python or C/C++
- Excellent communication and interpersonal skills, with the ability to effectively collaborate with customers and internal teams

Responsibilities:

- Model Tuning and Optimization: Spearhead efforts to optimize machine learning models for speed, memory efficiency, and accuracy through experimentation with different architectures, hyperparameters, and optimization techniques

Express Interest

Use the field below to send a message to this role's contact person. Your message is private: only that contact and the staffing team can see it. If you're in Tech, the hiring manager may request your GRAD from People Ops (see go/sharemygrad).

If you have questions about the role, but aren't yet ready to express interest, email the hiring manager listed to the left. If you're transferring into Google from another Bet, check out go/bet-mobility-faq first.

For more information regarding short term assignments, including eligibility, please visit go/transfers.

Have non-role specific questions about the Leadership Transfer Process at Google? Thinking about a Transfer with Level change? Curious about ways to optimize your application? Sign up for a confidential, 1:1 mobility session at go/leadershipmobility-oh or review curated resources at go/leadershipmobility .

To:

James Chamberlain
Mountain View
Recruiter, Leadership Staffing

Transfer message*

EXPRESS INTEREST

- Software and Hardware Optimization: Accelerate model training and inference by identifying and implementing software and hardware optimizations, which may include profiling code, optimizing data pipelines, and working with specialized accelerators (GPUs, TPUs, Tranium, etc)
- Framework Expertise: Showcase a strong understanding of deep learning frameworks such as PyTorch or JAX, including the ability to debug, extend, and optimize them
- Customer Collaboration: Translate customer requirements into technical solutions by working closely with them to understand their needs. This includes presenting technical findings and recommendations
- Performance Analysis: Identify bottlenecks and areas for improvement by developing and utilizing performance analysis tools

This role is eligible for a transfer with a level change (TLC). To learn about Googler eligibility and process requirements for applying and potentially transferring into this role from one level below, see go/transfer-tlc.

The US base salary range for this full-time position is $294,000-$414,000 + bonus + equity + benefits. Transfer compensation is determined algorithmically and is non-negotiable. Your recruiter will share more about the specific salary for your targeted location during the hiring process. Learn more about how a transfer may affect your compensation package , how location changes affect compensation, and about benefits at Google at go/benefits.

#jobcode#3411

Skills: Googleyness

○ My skills ○ Other skills