Full-Time Job

Principal Engineer, GPUs, Core ML, Borg, Spatial Flex - Sunnyvale

About this role

Full-Time Job (Level 8)

Location: Sunnyvale

Job family: Software Engineering

Job subfamily: Software Engineering

Product Area: TI Development

Updated Oct 15 | 341 views

Team members

Bill Jia
Sunnyvale
Vice President, Engineering ...

Jake Bercow
San Francisco
Executive Recruiter

Description

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.

Google's ML, Systems and Cloud AI (MSCA) organization builds the technical foundation behind Google's products as well as for Google Cloud's compute and AI/ML offerings. We manage the underlying design elements, developer platforms, product components, and infrastructure at Google. Core ML is the central machine learning platform team that provides ML software tools, services, solutions and infrastructure to all the Google product areas, including Search, Ads, Youtube, Cloud, Maps, etc. Core ML is focused on Driving ML Excellence for Google and the World. Our aim is to make it easier to perform ML experimentation, development and productionization and we work closely with Google Research and DeepMind to bring new ML models (e.g. Gemini) and innovation across the stack to market. This enables us to better meet the challenge of the rapidly evolving hardware and software space around ML.

Core ML is searching for a highly skilled and motivated Distinguished Engineer to play a leadership role in evaluating a collection of forthcoming hardware technologies for optimizing our offerings in this space. Advanced evaluation of GPU offerings and working closely with the internal TPU offerings to properly position and advocate for both cloud and internal customer GPU use-cases. This role will also directly contribute across Cloud and other product areas in creating the right infrastructure to access our ML Systems.

Minimum Qualifications:

- Bachelor's degree in Computer Science, a similar field, or equivalent practical experience.
- 15 years of experience in GPU performance related work
- 15 years of experience working with GPU Inference optimization

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: 

Jake Bercow
San Francisco
Executive Recruiter

Transfer message*

EXPRESS INTEREST

Preferred Qualifications:

- 20 years of professional experience in GPU performance related work at all levels of the stack
- Deep understanding of modern GPU architectures memory hierarchies, and performance bottlenecks
- Expertise in tailoring algorithms and ML models to exploit GPU strengths and minimize weaknesses
- Ability to develop and utilize sophisticated performance models and benchmarks to guide optimization efforts and hardware roadmap decisions
- Excellent communication and interpersonal skills, with the ability to effectively collaborate with customers and internal teams

Responsibilities:

- Providing engineering leadership and strategic product vision in this critical emerging Cloud use-cases as well as important internal use cases.
- 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.
- Customer Collaboration: Translate customer requirements into technical solutions by working closely with them to understand their needs. This includes presenting technical findings and recommendations.
- Mentorship and Leadership: Guide and inspire junior engineers by leading by example, sharing your expertise, and providing guidance on best practices.
- Performance Analysis: Identify bottlenecks and areas for improvement by developing and utilizing performance analysis tools.

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.

Skills: Software Engineering

My skills Other skills