DESCRIPTION

Are you excited about Machine Learning, chip acceleration, compilers, storage, systems or EC2? Are you passionate about delivering high quality services that affect hundreds of thousands of users? We are the dubbed the "secret sauce" behind AWS's success with development centers in the U.S. and Israel, Annarpuna is at the forefront of innovation by combining cloud scale with the world's most talented engineers.

The Annapurna team hires for multiple disciplines Software and Hardware engineers including but not limited to complier engineer, machine learning engineer, runtime engineer, performance engineer and ML chip accelerator, ASIC, physical designs, SDE in Test. Because of our teams' breadth of talent, we've been able to improve AWS cloud infrastructure in networking and security with products such as AWS Nitro, Enhanced Network Adapter (ENA), and Elastic Fabric Adapter (EFA), in compute with AWS Graviton and F1 EC2 Instances, in machine learning with AWS Neuron, Inferentia and Trainium ML Accelerators, and in storage with scalable NVMe.

Key job responsibilities
Innovating and delivering creative SW Designs to develop new services, solve operational problems, drive improvements in developer velocity, or positively impact operational safety
Writing requirements capturing documents, design documents, integration test plans, and deployment plans
Communicating status and progress of deliverables to schedule, and sharing learnings/ innovations with your team and stakeholders

BASIC QUALIFICATIONS

- Currently enrolled in, or completed a Bachelor's degree program or higher in Computer Science, Computer Engineering, Electrical Engineering or related field
- To qualify, applicants should have earned a Bachelor's or Master's degree between May 2023 to September 2025. Possible start dates for this role are between January 2025 to October 2025.
- Programming experience in internship or coursework with programming language such as Python and/or C or C++.

Candidates with strong interests and academic qualifications/research focus in two of the following:
- Knowledge of code generation, compute graph optimization, resource scheduling
- Data structure and algorithms
- Compiler - Optimizing compilers (internals of LLVM, clang, etc)
- Machine Learning - Experience with XLA, TVM, MLIR, LLVM
- Deep learning models and algorithms
- Tensorflow, PyTorch, or MxNET frameworks

PREFERRED QUALIFICATIONS

- Knowledge of code generation, compute graph optimization, resource scheduling
- Previous technical internship(s), if applicable.
- Experience in optimization mathematics such as linear programming and nonlinear optimization.
- Ability to effectively articulate technical challenges and solutions.

a disability and need a workplace accommodation or adjustment during the application and hiring process, including support for the interview or onboarding process, please visit https://amazon.jobs/content/en/how-we-hire/accommodations for more information. If the country/region you're applying in isn't listed, please contact your Recruiting Partner.

Our compensation reflects the cost of labor across several US geographic markets. The base pay for this position ranges from $99,500/year in our lowest geographic market up to $200,000/year in our highest geographic market. Pay is based on a number of factors including market location and may vary depending on job-related knowledge, skills, and experience. Amazon is a total compensation company. Dependent on the position offered, equity, sign-on payments, and other forms of compensation may be provided as part of a total compensation package, in addition to a full range of medical, financial, and/or other benefits. For more information, please visit https://www.aboutamazon.com/workplace/employee-benefits. This position will remain posted until filled. Applicants should apply via our internal or external career site.