Senior Software Engineer — AI/ML Systems Performance (US)

Company: Amazon (AWS Neuron)

Role family: SWE | Seniority: Senior | Location: United States (multiple locations)

Source (public posting): https://www.amazon.jobs/en/jobs/3117096/sr-software-engineer-ai-ml-aws-neuron-distributed-training

Cluster demand (ground-truth): {"SWE": 0.7, "MLE": 0.2, "DS": 0.05, "QD": 0.05, "QR": 0.0}

Reconstruction method: normalized sections + paraphrased responsibilities/qualifications for QA use.

ROLE SUMMARY
You will build and optimize software that enables large-scale model training and inference on specialized accelerators. The emphasis is systems engineering: performance tuning, tooling, and reliability for ML workloads.

KEY RESPONSIBILITIES
• Develop libraries, tools, and runtime components that support distributed training and inference.
• Profile and optimize performance for compute-heavy workloads (throughput, memory, kernel efficiency).
• Enable and tune modern model families (LLMs and vision models) on accelerator hardware.
• Collaborate with researchers and customer teams to diagnose issues and improve developer experience.
• Build automated tests and benchmarking suites to prevent regressions.

REQUIRED QUALIFICATIONS
• 5+ years building production software (systems or performance-oriented).
• Strong programming skills (C++/Python) and experience with profiling/performance analysis.
• Familiarity with ML training stacks and distributed compute concepts.

PREFERRED QUALIFICATIONS
• Experience with accelerator programming, compilers, or kernels.
• Experience with large-scale training frameworks and performance benchmarking.