Raj Kumar Thakur

LLM inference and ML systems on AWS Trainium

Cupertino, CA | rajanand.thakur@gmail.com | linkedin.com/in/raj-thakur-39756838 | rajthakur.dev | github.com/rajkthakur

Eleven years in software, the last four in AI/ML. I designed and built the disaggregated serving stack for vLLM on Trainium, started Neuron's move from PyTorch/XLA to native torch.compile, and now spend most of my time making models run faster and more accurately on Trainium.

EXPERIENCE

Senior Machine Learning Engineer

Dec 2022 – Present

Amazon Web Services, AWS Neuron (Annapurna ML), Cupertino, CA

Disaggregated serving

- Designed and built disaggregated inference for vLLM on Trainium. Prefill and decode run on separate nodes, the KV cache moves device to device with NIXL over EFA, and a batched runtime transfer API I added doubled KV throughput. It runs multi-node xPyD on EKS, including mixed GPU and Trainium.
- Led the work that removed the async gap between decode steps in disaggregated serving. The runtime was rebuilding DMA rings every step because inputs kept landing at new addresses.
- Built a multi-agent C++ code-review workflow on Claude Code that catches bugs in the team's runtime changes, including a concurrency race in async KV-offload code.

Performance and model bring-up

- Brought GPT-OSS-120B decode on native PyTorch to parity with XLA by fixing two compiler inputs, one of them a small all-gather mis-tagged as FSDP.
- Slow multi-LoRA was keeping NxDI customers from moving to vLLM Neuron. With the kernel's author, I designed the static-slot decode path for the NKI LoRA kernel, and vLLM Neuron then beat NxDI on multi-LoRA decode latency, which let us deprecate NxDI.
- Cleared a GPT-OSS-20B accuracy regression by isolating it to one compiler lowering, and validated the workaround.
- Brought up GPT-OSS on native PyTorch for Neuron, fixing integration bugs until BF16 passed GSM8K on Trainium2 and Trainium3. Tracked an out-of-memory failure to missing KV aliases and fixed the index bug that dropped them, bringing accuracy in line with XLA.

Native PyTorch for Neuron

- Started the native PyTorch project for Neuron inference (torch.compile lowered through torch-mlir instead of PyTorch/XLA). I proposed the native inference stack ahead of PyTorch Conference to put TorchNeuron and its ecosystem on stage there running natively, then wrote the proposal to retire PyTorch/XLA-based torch-neuronx in Neuron 2.32. All 8 required approvers signed off.
- Built device-free torch.compile for Neuron. I wrote and merged the compiler API (FX to StableHLO to NEFF) and a compile-only mode, so inference models compile on plain CPU hosts with no Trainium attached.
- Got the first vLLM TP8 server running on native PyTorch for Neuron, then dogfooded 6+ state-of-the-art models on it, including GPT-OSS and Qwen3-VL.
- Earlier: owned PyTorch upgrades across 10+ Neuron packages, moved the PJRT plugin to JAX 0.7.0, and built an HLO diff tool for compiler regressions.

Software Development Engineer II

Jun 2020 – Nov 2022

Amazon Web Services (Billing, OpenSearch), Bangalore, India

- Built a Step Functions service that generates and delivers billing documents for thousands of enterprise payers, which drove $300M in free-cash-flow improvements. Also built payer classification and mapping, which contributed to another $100M.
- Built a self-healing framework for OpenSearch nodes that cut customer root-cause tickets by 30%.

Senior Software Engineer, Backend

Jul 2018 – Jun 2020

Grab, Bangalore, India

- Rebuilt merchant settlement as an event-driven system on AWS Lambda (60% performance improvement). Designed the post-payment service in Go.

Software Engineer, Yatra Online; Software Developer, Tata Consultancy Services

Aug 2015 – Jul 2018

TECHNICAL SKILLS

LLM inference	vLLM, disaggregated prefill/decode, NIXL, EFA, tensor/data/expert parallelism, MoE, multi-LoRA, Trainium2/3
Compilers & runtimes	torch.compile, Dynamo, AOTAutograd, FX, StableHLO, PJRT, Neuron compiler and runtime, NKI
Debugging & perf	py-spy, gdb, system traces, roofline analysis, A/B benchmarking, accuracy debugging
Languages & tools	Python, C++, Go, Java, Docker, Kubernetes (EKS), Bazel

EDUCATION

B.Tech, Computer Science and Engineering, Jalpaiguri Government Engineering College (CGPA 8.5/10) 2015

TALKS & COMMUNITY

- Presenter, PyTorch Conference North America 2026: "Bringing High-Performance Inference Serving to Trainium with an Out-of-Tree vLLM Plugin"
- Amazon Bar Raiser (certified 2026), 100+ interviews. Mentor engineers across orgs.