

Muhammad Ibrahim

Senior Software Engineer · MLOps / LLM Inference Infrastructure

San Sebastián, Spain (remote, EU/UK hours) · +34 647 579 473 · ibrahimnd2000@gmail.com · mibrahim.dev ·
linkedin.com/in/ibrahimnd · github.com/ibrahimnd2000

PROFILE

MLOps engineer with 4+ years of experience who owns LLM serving infrastructure end to end: GPU kernels, KV-cache systems, routing, Kubernetes and the API layer. I led the serving stack for a frontier-scale MoE model (GLM-5.3) on multi-node NVIDIA B300 clusters for an enterprise customer, taking it from crashing and failing its benchmarks to passing its 200k-token SLA. I then re-architected it to serve **65% more throughput on 5 nodes instead of 11**, and shipped it to production with no downtime. I work from measurements (Nsight profiles, per-minute engine metrics, cost models and queueing maths), fix root causes in whichever layer they live, and contribute the fixes upstream to vLLM, LMCache and vLLM production-stack (~21 PRs with tests).

CORE SKILLS

LLM serving	vLLM, LMCache, prefill/decode disaggregation (NIXL over EFA RDMA), KV-cache offload, sharding and cache-aware routing; tensor, data, expert and context parallelism; MTP speculative decoding; FP8/NVFP4; CUDA graphs; FlashInfer
Performance	Nsight Systems/Compute, NCCL, py-spy, cost and queueing models, SLA benchmarking (TTFT, ITL, TPM, success rate), capacity sizing, A/B testing
Platform	Kubernetes, Docker, Helm, Argo CD, Karpenter, KEDA, Istio/Envoy; AWS (SageMaker HyperPod, EKS, EC2, EFA, ECR, Batch, Lambda, S3), Azure ML, GCP, Huawei Ascend (vLLM-Ascend, MindSpore); Terraform, CloudFormation, SkyPilot; GitLab CI, GitHub Actions, Jenkins; Prometheus/Grafana/Loki; Valkey/Redis
Languages	Python (asyncio, FastAPI, pytest, PyTorch), C++, Go, Rust, TypeScript/Node.js (Next.js, Vue/Nuxt), Java (Quarkus), SQL, GraphQL, Bash
AI engineering	Model compression and quantization (FP16/INT8/FP8/NVFP4), distributed training (DeepSpeed, Accelerate, torchrun, NeMo), agentic workflows, tool calling, RAG, LLM evaluation harnesses

EXPERIENCE

Mid/Senior Software Engineer (MLOps / LLMops) · Multiverse Computing Aug 2024 – Present · San Sebastián, Spain

LLM serving platform for an enterprise customer (GLM-5.3 FP8, up to 11 × 8-GPU B300 nodes)

- Owned the serving stack end to end** within 24 hours of joining (vLLM + LMCache on SageMaker HyperPod), taking it from crashing and failing its benchmarks to passing the 200k-token SLA: 64 concurrent users, 100% success, TTFT P50 13.1 s / P90 34.0 s against 30 s / 70 s limits, 7.7M tokens per minute on 11 nodes.
- Cut the hardware for the same SLA by more than half:** re-architected serving as prefill/decode disaggregation with context-parallel prefill, and passed the same SLA on **5 nodes at 12.7M tokens per minute**, 100% success. That's 65% more throughput than the 11-node deployment, 3.6× per node.
- Designed prefill/decode disaggregation** (vLLM NIXL connector over EFA RDMA): per-user decode speed went from 25.8 to 125–139 tokens/s on 200k-token prompts, with 0 failed KV transfers. Added P/D roles to the Helm chart and built the P/D router (cache- and load-aware prefill, data-parallel awareness, failover, circuit breaker, 80+ tests).
- Raised prefill throughput 34%** by profiling all 8 GPUs with Nsight Systems. Designed a sharded KV store across context-parallel ranks, with a correctness proof for its prefix lookup, and applied a FlashInfer FP8 kernel fix built for B300 (1.57 ms → 0.23 ms per call).
- Cut KV-cache tail latency 125×** (p99 84 s → 0.67 s on cache-hit-heavy traffic) without touching production. Replayed a production pod's exact request stream on an isolated node, used a cost model and queueing analysis to show a single-threaded Valkey L2 capped at 0.9 GB/s, then scaled it to one instance per GPU rank (6.3× bandwidth) and added asynchronous loading.
- Built KV-cache-aware routing** across replicas with safe fallbacks: 99% cache hit rate (up from 79–89%), ~8× cached-prompt throughput, warm-pass P50 9.8 s → 1.3 s. Fixed burst herding and in-flight-count leaks in the router's load signal (128-request burst 128/0/0 → 44/42/42). Enabled cross-node KV sharing over EFA (154–160 GB/s): a cached 188k-token prompt served from any replica in ~1.35 s instead of a 10.4 s prefill.
- Found and fixed silent correctness and stall bugs**, each validated on GPUs before rollout: KV corruption in LMCache (gibberish outputs from up to 16% to 0 of 2,000), a lost-completion race and a missing-key read hang in the native cache client, a mixed-KV-layout crash, and a load-failure stall (max ITL 575 s → 14.7 s).
- Reliability and operations:** zero-downtime production rollouts, including a 1.48× throughput upgrade and a dedicated customer P/D deployment: surge-first, router drain, image digests promoted unchanged. Root-caused truncated streams and 503s to autoscaler terminations and Istio/Envoy keep-alive mismatches in front of the model. Fixed native router crashes (patched libzmq), GPU-handoff races and Karpenter evictions. Built a per-minute SLA watcher.
- Open source (~21 upstream PRs with tests)** to vLLM, LMCache and vLLM production-stack: KV correctness and failed-load recovery, mixed KV layouts, a sharded KV store with async loading, per-rank remote cache instances, native client fixes, and

data-parallel and multi-model routing.

Agentic AI, customer delivery and automation

- **Agentic AI workshop for AWS (Madrid):** designed and delivered a hands-on workshop at the AWS office, showing a compressed LLM driving an agentic tool-calling workflow, with the whole environment as code (CloudFormation). Rated 4.6/5.
- **Customer deliveries:** deployed and validated a customer's full model-compression pipeline in 4 days; built automated IaC delivery pipelines for three environments with versioned releases, IAM user management and per-user usage dashboards. Automated model publishing (S3 → Hugging Face) and large dataset transfers (SharePoint ↔ S3).

LLM inference and GPU platform

- **7.5× LLM throughput** for Llama 3.3 70B on H200s (64 → 480 tokens/s; latency 5.49 s → 0.68 s) with CUDA graphs, speculative decoding, FlashAttention and FP8.
- Implemented mixed-precision and post-training quantization (FP16/INT8/FP8) for production LLMs in PyTorch, and designed Kubernetes GPU scheduling and autoscaling policies for LLM serving under strict p95 latency SLOs. Built internal observability for token-level latency and GPU utilization.
- Owned the company-wide GPU job platform (SkyPilot): multi-replica controllers with 0 failed jobs, EFA networking for a ~2× faster training pipeline, NVIDIA MIG partitioning as code. Led benchmarking of distributed training and inference stacks (DeepSpeed, Accelerate, torchrun, NeMo, vLLM). Deployed CompactifAI on Azure ML Studio and SageMaker HyperPod with cross-platform cost tracking.
- Architected Kubernetes infrastructure for quantum-classical workflows, cutting compute costs by 35% at 99.9% SLA compliance. Led the migration of legacy systems to modular Python/Next.js services: 70% fewer pipeline errors and 40% faster feature delivery.
- Championed the company's adoption of AWS Accelerate, securing **\$1.2M in contracts** through live demos. Mentored 4 junior engineers (2 promoted to mid-level within 6 months).

Junior Software Engineer · Multiverse Computing

Jun 2023 – Jul 2024 · San Sebastián, Spain

- **CompactifAI:** led a team of 6 building a 70%-compressed LLM, adopted by 3 enterprise clients for edge deployment ([arXiv:2401.14109](#)).
- Built an AWS Batch Monte Carlo simulation platform, cutting GPU costs by 25% (\$50K/year). Led a Streamlit threat-detection tool with 40% fewer false positives (\$120K lower breach-related losses).

Software Engineer · Bank Al Habib Ltd.

Sep 2022 – Jun 2023 · Karachi, Pakistan

- Automated 80% of loan-processing workflows with CI/CD pipelines (1,200+ loans/month), cutting processing time by 40%. Built secure Java/Quarkus microservices handling 50K+ daily transactions with zero security incidents.

Software Engineer Intern · Tapal Tea

Jun 2021 – Aug 2021 · Karachi, Pakistan

- Built an Android RFID warehouse app (30% less travel, 15+ labour hours saved weekly), automated SAP HANA workflows (40% fewer data-entry errors, 2 hours saved daily) and an Azure-integrated Power BI anomaly dashboard (+35% detection accuracy).

EDUCATION

MS Computer Science, Georgia Institute of Technology
BE Electrical & Computer Engineering, NED University of Engineering and Technology, Karachi · 2018 – 2022

LANGUAGES

English (native) · Spanish (B2)

CERTIFICATIONS

AWS Certified Machine Learning – Specialty (2025) · Microsoft Azure AI Engineer Associate (2025) · AWS Certified Developer – Associate · TensorFlow Certified Developer

VOLUNTEERING

AWS Certification Subject Matter Expert, GenAI (Arlington, VA, Jul 2025): designed, reviewed and validated exam content for the AWS GenAI Developer certification.