Position: Performance Engineer – AI Infrastructure & Database Systems
Company: Liqid, Inc
Work Location: Westminster, Colorado
(on-site only - this position is not remote)

Who We Are:

Liqid is doing for AI infrastructure what VMware did for x86 servers, bringing the agility of the cloud to the on-prem datacenter to dynamically allocate and share GPUs, memory, and other resources. This functionality is critical in running AI and advanced-data workloads in production (a.k.a. Inference). Liqid Matrix software enables global organizations to accelerate the adoption of AI, HPC and VDI by unlocking the bottleneck of GPU utilization and availability through dynamic pooling and sharing of AI resources for higher performance with less total cost, less power, and less hardware.

About the Role:

We are seeking an exceptionally strong early-career engineer for an internship or FTE role at the intersection of AI infrastructure, database systems, and performance engineering. This role is ideal for a highly intelligent candidate who has pursued database performance research in academia or at a major database vendor and wants to apply that background to cutting-edge AI system architecture.

The primary focus of this position is comparative KV-cache and memory-tier performance analysis across standard NVMe, HBM, server DRAM, and CXL-attached memory in modern AI infrastructure environments. The engineer will design and execute rigorous benchmarking methodologies, analyze workload behavior, and generate actionable insights that influence product architecture, system optimization, and customer-facing performance claims.

This role requires deep technical curiosity, strong experimental discipline, and the ability to work across AI inference infrastructure, storage and memory hierarchies, and multiple database paradigms including relational, vector, graph, and key-value systems.

LIQID

Key Responsibilities:

- Design and execute comparative performance tests for KV-cache behavior across NVMe, HBM, server DRAM, and CXL memory tiers.
- Develop benchmark methodologies for AI inference and data-intensive workloads, with emphasis on latency, throughput, tail behavior, scaling efficiency, and cost-performance tradeoffs.
- Build repeatable test harnesses, benchmarking frameworks, and data collection pipelines for system-level performance analysis.
- Evaluate workload characteristics relevant to LLM inference, retrieval, and memory/cache optimization.
- Analyze how different database architectures and storage/memory subsystems affect end-to-end application performance.
- Conduct performance studies spanning relational, vector, graph, and KV-cache-oriented data systems.
- Apply familiarity with benchmark environments and datasets such as STAC-M3, STAC-A2, FinanceBench, and LSEG/Refinitiv-style workflows and data environments.
- Produce clear technical reports, benchmark summaries, and recommendations for engineering, product, and executive audiences.
- Collaborate with hardware, software, systems, and product teams to guide infrastructure design decisions.
- Investigate bottlenecks across CPU, GPU, memory, interconnect, storage, and software stack layers.
- Help define credible, defensible performance claims for internal and external use.

Required Qualifications:

- MS/PhD (or MS/PhD track) in Computer Science, Computer Engineering, Electrical Engineering, or a related technical field. Exceptional BS candidates with highly relevant research experience may also be considered.
- Early-career engineer with a strong academic or industry background in database performance research.

- Research or practical experience at a leading university, research lab, or major database vendor.
- Strong understanding of database internals and performance behavior across multiple database types, including:
    - Relational databases
    - Vector databases
    - Graph databases
    - Key-value and cache-oriented systems
- Strong familiarity with benchmarking methodology, experimental design, and performance analysis.
- Knowledge of system architecture relevant to NVMe, HBM, server DRAM, and CXL.
- Experience profiling latency, throughput, IOPS, tail latency, memory bandwidth, cache behavior, and data movement bottlenecks.
- Programming ability in Python, C++, Java, or similar languages used in benchmarking and systems analysis.
- Must be self-sufficient at setting up hardware platforms involving PCI and CXL.
- Strong analytical and communication skills, including the ability to present results in a precise and defensible way.

Preferred Qualifications:

- Direct experience with AI inference systems, LLM serving, KV-cache optimization, or retrieval-augmented generation infrastructure.
- Familiarity with GPU-centric system design and heterogeneous memory architectures.
- Experience working with or evaluating financial-services benchmark workloads and environments.
- Familiarity with STAC-M3, STAC-A2, FinanceBench, and LSEG/Refinitiv-related datasets or performance use cases.
- Experience with performance tools, tracing, and profiling frameworks.
- Published academic research or patents in databases, systems, storage, caching, or performance engineering.
- Experience designing benchmarks that stand up to customer, partner, or third-party technical scrutiny.

LIQID

Ideal Candidate Profile:

The ideal candidate is intellectually rigorous, highly quantitative, and excited by hard performance questions that do not have obvious answers. They are comfortable moving between theory and hands-on experimentation, and they understand that strong benchmark results require careful workload design, reproducibility, and honest interpretation. They are equally capable of discussing database internals, memory hierarchy tradeoffs, and AI inference system behavior.

Compensation and Benefits:

- Base compensation: $135,000 - $175,000 / year depending on experience
- Generous Medical/Dental/Vision/Life/Disability benefits package
- 401K
- Unlimited PTO
- Cell phone stipend
- Free daily lunches provided in the office
- Flexible, casual work environment

Why Join Us:

At Liqid, you will have an opportunity to work on some of the most important infrastructure questions in AI: how memory, storage, and data system design shape real-world model performance. You will help define how next-generation AI platforms are measured, optimized, and differentiated.

Work Location Requirement:

This role is based in Westminster, CO and requires regular on-site presence. It is not a remote position.