

Vladimir Stryy

AI Infrastructure / Platform Lead — Lead SRE background

xeonerix.xyz github.com/XeoneriX t.me/xeonerix Moscow / remote

POSITIONING

Infrastructure / Platform Lead who operates private AI / GPU inference as production infrastructure, on a deep SRE base: private cloud, distributed storage, highload traffic platforms, DDoS resilience, observability and network telemetry. The differentiator is the fusion – an infrastructure engineer running LLM / RAG inference as an operated platform used by real engineering teams. Best-fit roles: AI / GPU Infrastructure Lead, Inference / AI Platform Lead, Head of Infrastructure, Principal / Staff SRE.

SUMMARY

8 years in SRE and infrastructure engineering. Strongest areas: Ceph and private cloud, highload infrastructure, P0/P1 incident response, multi-tenant observability, BGP/BMP telemetry, IaC, platform migrations and on-prem LLM/RAG systems for SRE workflows.

CORE SKILLS

- › **AI infrastructure:** private LLM serving, GPU inference, vLLM, llama.cpp, LiteLLM routing, Qdrant, bge-m3, reranking, RAG, Langfuse
- › **Private cloud:** Proxmox, OpenStack, VMware, Ceph, ZFS
- › **IaC:** OpenTofu, Terraform, Packer, Ansible, GitLab CI
- › **Observability:** VictoriaMetrics, VictoriaLogs, ClickHouse, Grafana, Vector
- › **Network:** BGP, BMP, SRv6, SR Policy, EVPN, FlowSpec, Juniper, Huawei
- › **Security & access:** Vault/OpenBao, RADIUS, Wazuh, audit trails
- › **Programming:** Go, Python, Bash, FastAPI, Flask, Celery
- › **Data & messaging:** PostgreSQL, Redis, NATS JetStream, ClickHouse

EXPERIENCE

Lead SRE / Infrastructure Lead

Cybersecurity / highload infrastructure platform · 2024–present

- › Lead infrastructure direction for a highload platform with carrier-scale L3/L4/L7 traffic and DDoS-mitigation workloads.
- › Build and operate private cloud on Proxmox, Ceph, OpenTofu, Packer, Ansible and GitLab CI.
- › Drive VM fleet migration from ESXi/vSAN-style to Proxmox/Ceph-style infrastructure with zero customer-visible downtime.
- › Design multi-tenant observability with VictoriaMetrics, VictoriaLogs, ClickHouse, Vector and Grafana HA.
- › Build BMP/BGP network telemetry, including SRv6, SR Policy, EVPN and FlowSpec.
- › Own architecture, migration planning, cost models and operational standards.
- › Provide technical leadership for a core team (2 direct reports) and set cross-team technical direction for ~5–7 network and adjacent infrastructure engineers by influence.
- › Build and operate a private on-prem LLM/RAG inference platform – LiteLLM routing, vLLM + llama.cpp GPU serving, Qdrant + bge-m3 retrieval with reranking, Langfuse tracing – for SRE automation, incident response and MR review; adopted across multiple engineering teams, first external adoption via an IDE coding-assistant (Cline) integration.

Senior SRE / Core Infrastructure SWAT

Large-scale consumer-internet / highload platform · 2022–2024

- › Core SRE/SWAT team handling critical infrastructure incidents and production degradations.
- › Escalation point for P0/P1 issues across storage, compute and platform infrastructure.
- › Stabilized Ceph clusters at ~100+ server and ~10 PB raw scale: disk failures, slow ops, recovery storms, rebalance, performance recovery.
- › Migrated Ceph services from Kubernetes-style placement to bare-metal-style operation to reduce operational complexity and improve reliability.
- › Contributed to Python services, CI/CD, PostgreSQL, Keycloak, Docker Swarm, object storage, performance tooling and observability.

Senior Cloud Infrastructure Engineer

International CDN / cloud provider · 2020–2022

- › Operated and developed distributed IaaS on OpenStack, Ceph and S3-compatible storage.
- › Supported 5 availability zones and ~80+ compute/storage servers; SLA for 50–80 B2B customers.
- › Built resource accounting and billing pipelines for VM, storage, IP and S3 usage with PostgreSQL.
- › Solved inconsistent S3 usage accounting with log-analysis pipelines.
- › Ceph expansion, balancing, recovery, Ansible bootstrap, capacity planning, hardware selection and customer migrations.

Systems Engineer / Python Developer

GDS / travel-tech · 2018–2020

- › Passenger-data processing pipelines replacing legacy shell workflows.
- › SFTP ingestion, normalization, validation, retry and delivery control for external systems, with audit trail.
- › Backend services in Python, Flask, Celery and PostgreSQL; query and index tuning.
- › Asterisk/LDAP telephony integration and automation bots.

PROJECTS

BGPeeK – open-source BGP looking glass

github.com/XeoneriX/bgpeek · Apache-2.0

- › Multi-vendor SSH (Juniper JunOS, Cisco IOS/XE/XR, Arista EOS, Huawei VRP), RPKI validation, parallel cross-device queries with diff.
- › RBAC (admin / NOC / public) with per-role output filtering; OIDC / LDAP / API-key auth; Fernet-encrypted SSH credentials; Redis rate limiting with circuit breaker; Prometheus /metrics; PostgreSQL audit log.
- › Stack: FastAPI, Jinja2, HTMX, Tailwind, PostgreSQL (asyncpg), Redis, Netmiko, Docker Compose.

Private AI / SRE inference platform

- › Operate a private LLM/RAG inference platform: LiteLLM routing, vLLM + llama.cpp GPU serving of Qwen-family models, Qdrant + bge-m3 retrieval with reranking, Langfuse tracing.
- › Treat private inference as an infrastructure workload – NUMA-aware placement, hugepages, CPU/memory isolation, concurrency limits, context-budget sizing.
- › Benchmarks for CPU-only isolation, GPU serving, prefix-cache behaviour and production co-tenancy impact.

EDUCATION

Moscow State Technical University of Civil Aviation – Information security of telecommunication systems