

Kostiantyn Dementiev

Systems researcher and platform engineer

Kyiv, Ukraine
✉ konstde00@gmail.com
📄 [konstde00](#)
ORCID [0009-0001-0795-1306](#)

Research interests

Systems for machine learning and the isolation they promise. I study isolation between tenants of shared systems: whether one tenant can read or infer another's data, how to test that separation and what it costs. The work began with database-per-tenant applications and now covers LLM serving, where tenants share a base model and its prefix cache but bring their own fine-tuned LoRA adapters. Next, I want to make correct isolation cheap enough to be the default. In my measurements, keying the cache on adapter content cost about 3% of throughput, compared with 66 to 79% for the fixes operators use first. More broadly, I am open to new research areas.

Education

Sep 2021–Jun 2025 **BSc in Computer Science**, *Taras Shevchenko National University of Kyiv (KNU)*, Faculty of Computer Science and Cybernetics
Program "Informatics", Department of Theory and Technology of Programming. One of the top 3 diploma marks in a group of 22 students. Selected courses (out of 100): Statistical Modeling in AI Problems (94), Distributed and Parallel Programming (91), Software Specification and Verification Methods (95), Cloud Computing (97).

Research experience

Jul 2025–present **Independent Researcher**, *unaffiliated, single-author work*
○ Extended the thesis into [2] to [4] and started a line of work on multi-tenant LLM serving [1].

Sep 2024–Jun 2025 **Undergraduate Researcher, bachelor's thesis**, *Faculty of Computer Science and Cybernetics, KNU*
Supervisor: Assoc. Prof. Oleksii M. Tkachenko, PhD. Thesis: "Methods for Building Distributed Multi-Tenant Applications with Spring Boot" (95 of 100).
○ Designed runtime onboarding (control-plane database, idempotent provisioning, atomically replaced routing table, broker-driven reconciliation): replicas gain a tenant and its database without a restart.
○ Built the prototype (Spring Boot, Kubernetes, PostgreSQL, Kafka) and measurement harness for [2] to [4].

Sep 2023–Jun 2024 **Undergraduate Researcher, term paper**, *Faculty of Computer Science and Cybernetics, KNU*
Supervisor: Asst. Prof. Anton O. Svystunov. Term paper graded 99 of 100.
○ Investigated runtime tenant onboarding for database-per-tenant Spring Boot applications, continuing my 2022 article and reference implementation.

Research manuscripts (single-author)

- [1] K. Dementiev. "Adapter Identity in Multi-Tenant LLM Serving: What a Name Cannot Separate." *Manuscript in preparation*.
Found a tenant-isolation failure on a shared GPU: vLLM keyed the prefix cache for a LoRA adapter by its client-supplied name, so tenants with same-named adapters shared cached blocks in 31 of 32 tested releases. Trained LoRA adapters with PEFT on Qwen2.5-0.5B-Instruct and served them with vLLM on an NVIDIA T4. An attacker identified a victim's fine-tuning domain nine times in ten among four candidates. Keying on a digest of the weights cost about 3% of throughput, against 66 to 79% for the usual fixes.
- [2] K. Dementiev. "Runtime Tenant Onboarding in Database-per-Tenant Spring Boot Applications: An Event-Driven Data-Source Routing Method for Distributed Deployments." *Under review at Information Technology and Implementation (IT&I 2026), Springer CCIS*.
Demonstrated on Kubernetes that every replica converged in all 160 provisionings, with a median convergence time of 419 ms at 3 replicas and 595 ms at 10.
Code and data for [2] to [4]: github.com/konstde00/runtime-tenant-onboarding
- [3] K. Dementiev. "A Verification Protocol for Silent Tenant-Isolation Failures in Dynamically Routed Database-per-Tenant Applications." *Under review at Advances in Cyber-Physical Systems*.
Designed a concurrent verification protocol that ties each isolation property to an observable that separates a pass from a silent failure. With the routing table replaced in place, 939 of 120,000 lookups reached the wrong tenant database, and none of 1,215,000 with atomic replacement. The protocol killed all 12 mutants, the existing tests one.
- [4] K. Dementiev. "Per-Tenant Connection Pools as TTL Caches: Capacity and Cold-Start Trade-offs in Database-per-Tenant PostgreSQL." *Under review at ACM SAC 2027*.
Modeled per-tenant connection pools as time-to-live caches. On minikube and Amazon EKS, the model predicted the mean of every pooler-free configuration to within 4%.

Experience

- 2025–present **Founder, *OpsServing, DevOps and cloud consultancy***
Scaled to about \$150k in annual recurring revenue in less than a year, across multiple client contracts and projects.
- Sep 2022–present **Software Engineer, backend and platform, *NCube Ltd. (contracted to ObservePoint)***
ObservePoint audits enterprise websites for analytics and privacy compliance. Its 26 data-processing and reporting services handle more than 100 TB a year. Joined as a backend developer and took over most of the platform work from late 2023.
- Co-designed consent-status resolution for the privacy-compliance product: trackers that matched no consent category were silently given the wrong status, and one shared SQL function now reports them as uncategorized.
 - Prototyped cross-customer isolation for the company's GitHub App: an installation is accepted only if the organization owns it and the user administers it. Replaced long-lived CI keys with OIDC roles across about 30 repositories and assisted with security hardening for the company's SOC 2 certification.
 - Traced recurring Step Functions failures in a production data-deletion pipeline to memory-starved containers that let task tokens expire, and rolled the fix out to 9 services.
 - Diagnosed why tripling an enterprise tenant's concurrency limit (10 to 30) left its throughput flat: the scheduler admitted one audit per tenant per cycle, so only about 10 ran at a time. Proposed a bounded per-tenant admission rate.
 - Benchmarked four ways to read 2.1 million small S3 files. Compacting them with Athena before loading into Snowflake let a backfill that had failed after 4 hours finish in about 6 minutes. A K-means model of object sizes showed that deleting early beat lifecycle expiry for 97% of files.
 - Traced a tenfold latency spike at double the request rate to connection-pool starvation and fixed it, cutting the endpoint's CPU use by 34%. Separately, cut an enterprise report page's load time from 17s to under 2s and a 10-table analytics query over about 10 TB from 21s to 2.1s.
 - Led the platform modernization: 11 services moved from Rancher to ECS and Fargate, 11 Jenkins pipelines to CircleCI, Consul retired across 8 services, and Redis moved to ElastiCache without downtime.
 - Led weekly incident reviews for the reporting services (2023 to present), wrote about 20 design documents and mentored two junior developers.
- Nov 2021–Sep 2022 **Java Developer, *Impltech***
- Built backend features for marketing-automation and language-learning products starting in the first year of my BSc, and set up the team's CI deployments to AWS and its integration tests.

Honors

- 2021 Student Mathematics Olympiad, Faculty of Mechanics and Mathematics, KNU: 1st place, 50 of 50 points
- 2023–2026 Operational Excellence Champion, ObservePoint: 3 of 13 quarterly awards, the most of any recipient
- 2025 Product and Engineering Team Member of the Year 2025 (Ukraine), ObservePoint
- 2024 Take Accountability award, ObservePoint, for the Rancher, Jenkins and Nexus migrations
- 2023 ObservePoint AI Hackathon, "Most Impactful" winner for an AI grammar and spelling checker

Extracurricular activities and community service

- Writing Medium: five technical articles, 2022 to 2025, about 20,000 views; DOU, Ukraine's largest developer community: one article, 2024, 3,000+ views
- Running Started a running club with friends in 2025, and it now has 7 members
- Volunteering DevOps for Charity Hero, a project distributing humanitarian aid in eastern Ukraine

Certifications

CKA, CKAD, CKS; AWS Security Specialty; AWS Solutions Architect, Developer and SysOps Administrator Associate; HashiCorp Terraform Associate; Snowflake SnowPro Core

Open source

- Pie LLM serving: sessions shared one name, so any client could control other users' processes (#1101, #1100).
- SREGym AI-agent incident benchmark: three grading oracles accepted wrong fixes (#1107, #1108, #1109).
- ZMap 6.4k stars: a BACnet or DNS reply could end a scan, and option parsing overran packets (#984, #983).
- PyHealth 1.7k stars: its cleanup deleted other jobs' temp folders on a shared cache (#1296).
- NTP Pool CCS 2026 paper artifacts: the IPv6 traceroute never probed the last hop limit (#1).
- Redash 28.8k stars: Google sign-in behind load balancers that terminate TLS (#7178, 2025).

Skills

- Programming Python, Java, Scala, SQL, Bash
- ML systems PyTorch, Hugging Face Transformers and PEFT (LoRA fine-tuning), vLLM (multi-LoRA serving)
- Data Snowflake, Kafka, DuckDB, PostgreSQL, Amazon Athena
- Infrastructure AWS, Kubernetes, Terraform, Docker, CircleCI, GitHub Actions
- Languages English (C1, IELTS 8.0, 2025), Ukrainian (native)