Alexey Shvechkov

Greater Boston, MA | alexey@shvechkov.com | +1.774.278.1743 | [LinkedIn profile] | [github.com]

Summary:

Technical Director | Hands-On Systems, Cloud & AI Architect | From Kernel Drivers to Agentic AI Platforms

Hands-on Technical Director with over 20 years of expertise in architecting scalable systems, cloud-native solutions (AWS, Kubernetes), and AI platforms (LLMs, RAG). Lead global teams while coding mission-critical features, reducing delivery cycles by up to 40% and driving innovation in high-throughput environments.

Education:

- M.Sc. in Applied Mathematics – Institute of Telecommunications and Computer Science, Russia (1996)
- Coursera Machine Learning Specialization, Stanford Online 2024 - Certificate
- Holder of multiple U.S. patents in storage and infrastructure technologies

Skills:

Systems Programming

- C++, Golang, Rust, Python; Linux/Windows kernel drivers (storage stack, eBPF, KMDF/WDM, minifilters, file systems)

Distributed Systems & Cloud

- AWS (EC2, S3, Lambda/serverless, API Gateway), Azure, Kubernetes, Docker, Terraform
- Scalable microservices, fault-tolerant architectures, CI/CD pipelines

AI/ML & LLMs

- PyTorch, TensorFlow, Hugging Face, Unsloth, LangChain; RAG, transformer fine-tuning (full, LoRA), distributed training, hosting models locally; Experience with AWS AI/ML Pipeline tools (Glue Catalog, ETL, DataBrew, Athena, Sagemaker,)

Databases

- RDBMS (PostgreSQL, MySQL), NoSQL (Redis, Cassandra)

Leadership

- Technical direction, agile execution, mentoring teams of up to 10 engineers

Additional Skills

- React, Next.js, scripting, Git, RESTful API design, performance tuning, system-level debugging, rapid prototyping

Work Experience:

Technical Director, Software Engineering (Arcserve – *Greater Boston, MA | 2014 – Present)*

Progressed from Staff SWE to Sr Architect to Product Owner to Technical Director

Leadership & Strategy

- Provided hands-on technical leadership for AI/ML initiatives and platform modernization, reporting to CTO; Owned end-to-end lifecycle of three enterprise products (Replication & High Availability, OneXafe, CloudDirect, $15-17M revenue).
- Led multiple high-impact teams of up to 10 engineers across storage, cloud protection, and AI-driven features, defining strategic roadmaps, architecting scalable solutions, and delivering on schedule with up to 40% reduced delivery cycles.
- Mentor engineers while actively coding and prototyping, ensuring technical excellence and measurable customer impact.

AI/ML Integration & Agentic Systems

- Architected and prototyped malware/anomaly detection features: coded EDA pipelines, performed feature engineering, trained ML models, and fine-tuned LLMs using Python and C++.
- Developed and optimized transformer-based masked LLMs via ONNX Runtime for edge inference, writing performance-critical C++ code.
- Prototyped and coded agentic AI assistants using LangChain and OpenAI, integrating with legacy systems via REST, SOAP, and MCP protocols. (RAG knowledgebase chat bots + agentic backend features – decreased support calls by 20%)
- Designed and implemented an interactive RAG-based AI assistant, writing core logic and optimizing retrieval pipelines.
- Architected and coded microservices (Model Context Protocol servers) to integrate AI-driven interfaces with legacy APIs.

Systems & Product Engineering

- Architected and coded core Arcserve features, including data deduplication, file system/server replication, high availability (HA), and agentless protection for AWS/Azure workloads (VMs, containers, storage), with hands-on contributions in C++ and Go.
- Prototyped and developed a Linux-based immutable object-store server (cyber-resilient appliance), writing core components and Reduced delivery cycle for immutable object-store server by 40%, enabling 6-month faster market entry.
- Coded system and kernel-level components (file systems, filter drivers) on Windows and Linux for high-throughput backup and replication products.

R&D & Innovation

- Won hackathons for AI and infrastructure prototypes, rapidly coding proof-of-concepts in Rust and Python.
- Led M&A-influencing projects, performing due diligence and coding prototypes to validate technical feasibility.
- Developed kernel-mode anti-malware solutions and email archiving engines, contributing both architecture and implementation.

Early Career:

Software Engineer / Principal Engineer (*Yandex, XOsoft, Computer Associates | 1998 – 2014)*

- Designed and coded high-traffic web applications and CDN modules at Yandex and XOsoft using C++ and Perl, optimizing search, advertising, and content delivery systems for performance and reliability.
- Built high-availability data replication modules at XOsoft using C++ and led a team of 5 engineers, later architecting features like deduplication, full system protection, and cloud replication while developing file system drivers for Windows and Linux/Unix at CA to deliver scalable solutions.
- Developed an end-to-end build automation system akin to Jenkins for Unix/Linux platforms, integrating a web UI for progress reporting, reducing manual labor by 40–50%, and streamlining QA and production processes.

Open-source projects/POCs (highlights)

- s3stor – deduplicating archiving /backups into S3 compatible storage (Golang)
- gos3rve – exposing file local systems via s3 APIs (Golang)
- ufc – fast unique file copy/ indexes and stores unique files (rust)