Dima (Dan) Sivov

SOFTWARE ARCHITECT · APPLIED-AI RESEARCHER

Katzir, Israel · +972-52-868-4276 · dsivov@gmail.com · github.com/dsivov

Software architect and researcher who takes ideas from concept to validated prototype fast, then leads teams to turn the winners into shipped products and large-scale research. Currently Principal Engineer at Toga Networks (Huawei), working across semantic search, visual language models, RAG, and agentic AI, with peer-reviewed publications and open-source contributions to OpenStack, openGauss, and RAMCloud.

EXPERIENCE

Principal Engineer & Researcher, **Toga Networks (Huawei)** 2013 – Present

- Lead applied research in information retrieval, semantic search, and large-scale visual deep-understanding — carrying ideas from prototype through to production systems and peer-reviewed publication.
- Architect and build distributed database and networking components adopted into Huawei platforms.
- Prototype new AI approaches — RAG, agentic systems, visual language models — and lead engineering teams that move the strongest into product and large research programs.

Team Leader, Event-Driven Monitoring & Management Systems, **BetterPlace** 2010 – 2013

- Led the team designing and building event-driven monitoring and management systems for a large electric-vehicle charging network.

R&D Manager · Project Manager · System Engineer, **Matrix (Technology Division)** 2002 – 2010

- Led R&D teams and end-to-end delivery of enterprise software projects across multiple clients.

Earlier — **Intel** (Israel Development Center), Unix infrastructure software · **Philips Medical Systems** (Elscint / Picker), system integration & connectivity 1995 – 2002

Independent Consultant — software architecture, system design & technology strategy for startups Ongoing

SKILLS

Languages Python, C, C++, Java

Focus Distributed systems · Databases · Information retrieval & semantic search · RAG · Deep learning · LLMs & VLMs · Agentic AI · System integration

INDEPENDENT RESEARCH

Speculative Context: A Pool-Based Cache for SOP-Constrained Voice Agents ***(2026, sole author)***
A two-system architecture that hides external-data latency in real-time voice agents using a speculatively-populated pool cache and an LLM-free reranker.

PUBLICATIONS

- **Few-Shot Text-to-Image Retrieval: New Benchmarking Dataset and Optimization Methods** — *arXiv, 2026*
- **NSFL: A Post-Training Neuro-Symbolic Fuzzy Logic Framework for Boolean Operators in Neural Embeddings** — *arXiv, 2026*
- **Does the Performance of Text-to-Image Retrieval Models Generalize Beyond Captions-as-a-Query?** — *ECIR 2024*
- **Industrial-strength OLTP using main memory and many cores**
- **Fiber-based architecture for NFV cloud databases** — *Proc. VLDB Endowment, 2017*

Full list on ACM Digital Library →

OPEN SOURCE

Context Graph ***(sole author · Python, FastAPI)***
A decision-aware knowledge-graph system that extends LightRAG's graph-RAG with contextual quadruples (h, r, t, rc), so every edge carries its full decision lineage.

investigator ***(sole author · Python, Svelte)***
OSINT pipeline that turns a news corpus into a scored, fully source-traceable knowledge graph.

openGauss
Architect for Memory-Optimized Tables (MOT).

OpenStack
Database support for Open vSwitch.

RAMCloud
DPDK network support.

EDUCATION

University of Haifa — Mathematics & Computer Science
Moscow University of Communications — Television & Broadcasting